

**Titre:** Prévention des attaques par logiciels malveillants: perspectives de  
Title: la santé publique

**Auteur:** Fanny Lalonde Lévesque  
Author:

**Date:** 2018

**Type:** Mémoire ou thèse / Dissertation or Thesis

**Référence:** Lalonde Lévesque, F. (2018). Prévention des attaques par logiciels malveillants:  
Citation: perspectives de la santé publique [Ph.D. thesis, École Polytechnique de Montréal].  
PolyPublie. <https://publications.polymtl.ca/3315/>

 **Document en libre accès dans PolyPublie**  
Open Access document in PolyPublie

**URL de PolyPublie:** <https://publications.polymtl.ca/3315/>  
PolyPublie URL:

**Directeurs de  
recherche:** Jose Manuel Fernandez  
Advisors:

**Programme:** Génie informatique  
Program:

UNIVERSITÉ DE MONTRÉAL

PRÉVENTION DES ATTAQUES PAR LOGICIELS MALVEILLANTS: PERSPECTIVES  
DE LA SANTÉ PUBLIQUE

FANNY LALONDE LÉVESQUE  
DÉPARTEMENT DE GÉNIE INFORMATIQUE ET GÉNIE LOGICIEL  
ÉCOLE POLYTECHNIQUE DE MONTRÉAL

THÈSE PRÉSENTÉE EN VUE DE L'OBTENTION  
DU DIPLÔME DE PHILOSOPHIÆ DOCTOR  
(GÉNIE INFORMATIQUE)  
SEPTEMBRE 2018

UNIVERSITÉ DE MONTRÉAL

ÉCOLE POLYTECHNIQUE DE MONTRÉAL

Cette thèse intitulée :

PRÉVENTION DES ATTAQUES PAR LOGICIELS MALVEILLANTS: PERSPECTIVES  
DE LA SANTÉ PUBLIQUE

présentée par : LALONDE LÉVESQUE Fanny  
en vue de l'obtention du diplôme de : Philosophiæ Doctor  
a été dûment acceptée par le jury d'examen constitué de :

M. MERLO Ettore, Ph. D., président

M. FERNANDEZ José M., Ph. D., membre et directeur de recherche

M. DUPONT Benoît, Doctorat, membre

M. DACIER Marc, Doctorat, membre externe

## DÉDICACE

*À mon petit Tristan,  
et à tous ceux qui ont cru en moi.*

## ÉPIGRAPHE

*They did not know it was impossible so they did it.*

- Mark Twain

## REMERCIEMENTS

Les remerciements qui suivent s'adressent à tous ceux qui, de près ou de loin, ont contribué de par leur support, leur gentillesse, leur amour, et leur temps à cette aventure de quatre ans.

J'aimerais en un premier temps remercier les membres du jury pour leur temps et leurs contributions à cette thèse. Je tiens particulièrement à adresser mes remerciements à mon directeur de recherche, et ami, José M. Fernandez, pour m'avoir accompagnée et encouragée tout au long de mon projet. Mes remerciements iront également aux commanditaires de la thèse : la Corporation Microsoft, le Fonds de Recherche du Québec –Nature et Technologies, ainsi que le Conseil de recherches en sciences naturelles et en génie du Canada.

Je souhaiterais de plus remercier toutes ces personnes qui ont contribué à la réussite de ce travail. Je ne peux que remercier Dennis Batchelder, aujourd'hui chez AppEsteem et directeur chez Microsoft au moment où je débute ma thèse. Dennis : *Thank you for everything. It is always a pleasure to collaborate with you. You are a great source of inspiration and motivation.* Anil Somayaji et Sonia Chiasson, de l'Université Carleton, je vous remercie pour votre temps, vos idées, votre patience, et votre support. Sur une note plus personnelle, je tiens à remercier ma famille et mes ami(e)s pour leur patience et leur support ; mon Amour Jessy Campos, qui a su être là pour moi ; et finalement Tristan Yiddir Lalonde, mon petit homme, pour avoir ensoleillé mes journées.

J'aimerais de plus profiter de ces quelques lignes pour remercier les étudiants et les membres du personnel de l'université avec qui j'ai été amenée à travailler au cours des 12 dernières années. Mon passage à Polytechnique a été pour moi une expérience extraordinaire, et je vous en remercie. Finalement, je tiens à adresser mes remerciements à Element AI, plus particulièrement à mes collègues de l'équipe sécurité, pour leur support et leurs encouragements.

## RÉSUMÉ

L'augmentation de la connectivité et du développement des infrastructures numériques a contribué à multiplier les motivations et les opportunités des attaques informatiques. Bien que plusieurs progrès aient été réalisés au niveau du développement et de l'implémentation de stratégies de protection, la majorité de ces efforts sont dédiés au développement de nouvelles solutions, et non à leur évaluation et leur promotion. Il devient dès lors essentiel pour les gouvernements, les entreprises, et les individus de définir des modèles et des moyens de coopération permettant d'identifier et d'évaluer les stratégies visant à réduire le risque que posent les menaces informatiques.

À cet effet, le domaine de la sécurité des systèmes d'information pourrait bénéficier des leçons apprises et des méthodes utilisées dans le domaine de la santé. En particulier, nous croyons que l'adoption d'une perspective axée sur l'approche de la *santé publique* permettrait de fournir un cadre global pour i) identifier les facteurs qui affectent la sécurité des systèmes d'information et en comprendre les causes sous-jacentes, ii) développer et évaluer des stratégies efficaces visant à améliorer la sécurité des systèmes d'information, et iii) implémenter et disséminer auprès de la population les stratégies développées.

Dans le cadre de la présente thèse, nous proposons de nous inspirer des méthodes en santé publique pour développer un modèle de prévention applicable au contexte des attaques par logiciels malveillants. Notamment, nous appliquons notre modèle de prévention afin d'identifier les causes et les corrélats reliés aux attaques par logiciels malveillants, et d'évaluer l'efficacité réelle des solutions antivirus à prévenir ces attaques. À partir de données réelles d'attaques par logiciels malveillants, nous avons réalisé cinq études empiriques ; trois visant à identifier des facteurs de risque et des facteurs de protection, et deux visant à évaluer l'efficacité des antivirus dans un environnement réel.

Les résultats de nos travaux de recherche ont, entre autres, permis : i) d'identifier de nouveaux facteurs de risque et de protection reliés aux attaques par logiciels malveillants, ii) d'identifier des sous-populations à risque plus élevé, et iii) de mettre en évidence comment l'effet des facteurs identifiés et des solutions antivirus varie selon le contexte (type de menace, environnement, usager, etc.). Qui plus est, la présente thèse a permis de valider la viabilité et le potentiel d'une approche basée sur la santé publique en sécurité des systèmes d'information.

## ABSTRACT

The increased connectivity and development of digital infrastructures has yielded to increased motivation and opportunities for computer threats. Although there has been some progress in the development and implementation of protection strategies, the majority of these efforts are dedicated to the development of new solutions, and not to their evaluation and promotion. It is therefore essential for governments, businesses, and individuals to develop models and means of cooperation in order to identify and evaluate effective strategies aimed at reducing the risk posed by computer threats.

To this end, the field of information security could benefit from lessons learned and methods used in health. In particular, we believe that adopting a *public health* perspective could provide a comprehensive framework for i) identifying and understanding the factors that affect the information systems security and understand their underlying causes, ii) develop and evaluate effective strategies to improve the security of information systems, and iii) implement and disseminate the strategies developed to the population.

In this thesis, we propose to use public health methods to develop a prevention model for the context of malware attacks. In particular, we apply our prevention model to identify the causes and correlates of malware attacks, and evaluate the effectiveness of antivirus solutions in preventing computer threats. Using real-world malware attacks data, we conducted five empirical studies ; three to identify risk factors and protective factors, and two to assess the effectiveness of antivirus in a real-world environment.

The results of our research allowed us, among others, to : i) identify new risk and protective factors related to malware attacks, ii) identify high-risk sub-populations, and iii) highlight how the effect of the identified factors and antivirus solutions vary by context (type of threat, environment, user, etc.). In addition, this thesis validated the viability and potential of a public health approach to information security.



## TABLE DES MATIÈRES

DÉDICACE . . . . .	iii
ÉPIGRAPHE . . . . .	iv
REMERCIEMENTS . . . . .	v
RÉSUMÉ . . . . .	vi
ABSTRACT . . . . .	vii
TABLE DES MATIÈRES . . . . .	viii
LISTE DES TABLEAUX . . . . .	xiii
LISTE DES FIGURES . . . . .	xvi
LISTE DES SIGLES ET ABRÉVIATIONS . . . . .	xvii
LISTE DES ANNEXES . . . . .	xviii
CHAPITRE 1 INTRODUCTION . . . . .	1
1.1 Éléments de la problématique . . . . .	1
1.2 Motivations . . . . .	3
1.3 Objectifs de recherche . . . . .	4
1.4 Contributions . . . . .	5
1.5 Publications . . . . .	6
1.6 Plan de la thèse . . . . .	7
CHAPITRE 2 SÉCURITÉ ET SANTÉ PUBLIQUE . . . . .	9
2.1 Logiciels malveillants . . . . .	9
2.1.1 Classification . . . . .	9
2.1.2 Moyens de propagation . . . . .	11
2.1.3 Acteurs malveillants et motivations . . . . .	12
2.1.4 Tendances et prévalence des logiciels malveillants . . . . .	14
2.2 Logiciels antivirus . . . . .	16
2.2.1 Technologies et méthodes de détection . . . . .	17

2.2.2	Marché et produits antivirus . . . . .	18
2.3	Santé publique . . . . .	19
2.3.1	Définitions et historique . . . . .	19
2.3.2	Déterminants de la santé . . . . .	20
2.3.3	Fonctions essentielles . . . . .	23
2.3.4	Cadre de travail . . . . .	23
2.3.5	Modèles d'implémentation . . . . .	25
2.4	Sécurité et santé publique : Travaux antérieurs . . . . .	26
2.4.1	Travaux de recherche . . . . .	26
2.4.2	Implémentation et initiatives existantes . . . . .	28
CHAPITRE 3	MODÈLE DE PRÉVENTION . . . . .	29
3.1	Cadre de travail . . . . .	29
3.1.1	Définition du problème . . . . .	30
3.1.2	Identification des déterminants . . . . .	30
3.1.3	Développement et évaluation de stratégies . . . . .	31
3.1.4	Implémentation et promotion des stratégies . . . . .	31
3.2	Application aux attaques par logiciels malveillants . . . . .	32
3.2.1	Déterminants des attaques par logiciels malveillants . . . . .	32
3.2.2	Efficacité des logiciels antivirus . . . . .	40
3.2.3	Source de données . . . . .	46
CHAPITRE 4	ARTICLE 1 : NATIONAL-LEVEL RISK ASSESSMENT : A MULTI-COUNTRY STUDY OF MALWARE INFECTIONS . . . . .	48
4.1	Introduction . . . . .	48
4.2	Previous studies . . . . .	51
4.3	Study design and methods . . . . .	53
4.3.1	Data collection . . . . .	54
4.3.2	Statistical methods . . . . .	57
4.4	Results . . . . .	59
4.4.1	Global model . . . . .	60
4.4.2	Model by socio-economic status . . . . .	63
4.5	Interpretation . . . . .	66
4.6	Study limitations . . . . .	70
4.7	Conclusion and policy implications . . . . .	72
CHAPITRE 5	ARTICLE 2 : TECHNOLOGICAL AND HUMAN FACTORS OF MAL-	

WARE ATTACKS : A COMPUTER SECURITY CLINICAL TRIAL APPROACH	75
5.1 Introduction . . . . .	76
5.2 Related work . . . . .	78
5.2.1 Antivirus product evaluation . . . . .	78
5.2.2 Human factors and computer threats . . . . .	80
5.3 Computer security clinical trials . . . . .	83
5.4 Study description . . . . .	84
5.4.1 Ethics clearance . . . . .	84
5.4.2 Equipment . . . . .	85
5.4.3 Experimental protocol . . . . .	86
5.5 Antivirus evaluation . . . . .	89
5.5.1 Threats detected by antivirus . . . . .	89
5.5.2 Missed threats . . . . .	90
5.5.3 Antivirus efficacy . . . . .	92
5.5.4 User experience . . . . .	92
5.6 User Profiling and Behaviour . . . . .	97
5.6.1 Characteristics and demographic factors . . . . .	97
5.6.2 Behavioural factors . . . . .	102
5.7 Study limitations . . . . .	111
5.8 Discussion and conclusion . . . . .	112
CHAPITRE 6 ARTICLE 3 : AGE AND GENDER AS INDEPENDENT RISK FAC-	
TORS FOR MALWARE VICTIMISATION . . . . .	115
6.1 Introduction . . . . .	115
6.2 Previous studies . . . . .	117
6.2.1 Demographics and malware victimisation . . . . .	117
6.2.2 Demographics and other computer threats . . . . .	118
6.3 Study design and methods . . . . .	118
6.3.1 Case-control study design . . . . .	119
6.3.2 Target population . . . . .	119
6.3.3 Data collection . . . . .	120
6.3.4 Ethical and privacy considerations . . . . .	120
6.3.5 Statistical analysis . . . . .	121
6.4 Results . . . . .	121
6.4.1 Population . . . . .	121
6.4.2 Malware encounter risk factors . . . . .	122

6.4.3	Risk factors by malware types . . . . .	125
6.5	Discussion . . . . .	126
6.5.1	Gender difference . . . . .	126
6.5.2	Age difference . . . . .	129
6.5.3	Summary of findings . . . . .	131
6.6	Study limitations . . . . .	132
6.7	Conclusion . . . . .	133

## CHAPITRE 7 ARTICLE 4 : MEASURING THE HEALTH OF ANTIVIRUS ECOSYS-

TEMS . . . . .	136
7.1 Introduction . . . . .	136
7.2 Background . . . . .	137
7.3 Antivirus ecosystem indicators . . . . .	139
7.3.1 Activity . . . . .	139
7.3.2 Diversity . . . . .	140
7.3.3 Stability . . . . .	141
7.4 Country level analysis and evaluation . . . . .	143
7.4.1 Activity . . . . .	147
7.4.2 Diversity . . . . .	149
7.4.3 Stability . . . . .	149
7.5 Discussion . . . . .	150
7.6 Conclusion . . . . .	151

## CHAPITRE 8 ARTICLE 5 : ARE THEY REAL ? REAL-LIFE COMPARATIVE TESTS OF ANTI-VIRUS PRODUCTS . . . . .

8.1	Introduction . . . . .	153
8.2	Real-life anti-virus evaluation . . . . .	154
8.3	Study design and methods . . . . .	155
8.3.1	Cohort study design . . . . .	155
8.3.2	Study population . . . . .	157
8.3.3	Data collection . . . . .	157
8.3.4	Exposure (AV protection) . . . . .	158
8.3.5	Statistical analysis . . . . .	158
8.4	Results . . . . .	160
8.4.1	Anti-virus effectiveness . . . . .	160
8.4.2	Anti-virus effectiveness by vendors . . . . .	164
8.5	Study limitations . . . . .	169

8.6 Conclusion . . . . .	170
CHAPITRE 9 DISCUSSION GÉNÉRALE . . . . .	173
9.1 Modèle de prévention des attaques par logiciels malveillants . . . . .	173
9.1.1 Identification des déterminants . . . . .	173
9.1.2 Évaluation de stratégie . . . . .	174
9.2 Contributions et implications . . . . .	175
9.3 Sécurité des systèmes d'information publique . . . . .	176
CHAPITRE 10 CONCLUSION . . . . .	178
10.1 Synthèse des travaux . . . . .	178
10.1.1 Modèle de prévention . . . . .	178
10.1.2 Identification des déterminants . . . . .	178
10.1.3 Évaluation de stratégie . . . . .	180
10.2 Limitations et travaux futurs . . . . .	181
10.3 Conclusion . . . . .	183
RÉFÉRENCES . . . . .	184
ANNEXES . . . . .	199

## LISTE DES TABLEAUX

Tableau 3.1	Modèle de prévention des attaques par logiciels malveillants . . . . .	29
Tableau 3.2	Indicateurs d'attaques, population et déterminants étudiés (niveau institutionnel et organisationnel) . . . . .	34
Tableau 3.3	Indicateurs d'attaques, population et déterminants étudiés (niveau individuel) . . . . .	35
Table 4.1	Country-level factors . . . . .	55
Table 4.2	Descriptive statistics (whithout transformation) . . . . .	58
Table 4.3	Descriptive statistics after log-transformation . . . . .	58
Table 4.4	Pearson correlation coefficients between infection rate and country-level factors . . . . .	60
Table 4.5	Global multiple general linear regression results (N=50 countries) . .	61
Table 4.6	Pearson correlation coefficient between infection rate and country-level factors . . . . .	64
Table 5.1	User experience and opinion per month . . . . .	94
Table 5.2	Proportion of users for each factor . . . . .	98
Table 5.3	Odds ratio of user characteristics and demographic factors . . . . .	98
Table 5.4	Odds ratio of behavioural factors . . . . .	102
Table 5.5	Type of applications installed per month . . . . .	103
Table 5.6	Type of applications installed by others per month . . . . .	104
Table 5.7	Most frequently used applications per month . . . . .	104
Table 5.8	Second most frequently used applications per month . . . . .	104
Table 5.9	Primary location from which the laptop was connected to the Internet per month . . . . .	105
Table 5.10	Installed web browsers . . . . .	106
Table 5.11	Most frequently used web browser . . . . .	106
Table 5.12	Security and privacy default settings . . . . .	106
Table 5.13	Odds ratio by web page categories . . . . .	108
Table 5.14	Odds ratio by type of files downloaded . . . . .	109
Table 5.15	Computer security expertise . . . . .	110
Table 5.16	Concern about the computer's security per month . . . . .	110
Table 6.1	Population demographics by factor . . . . .	121
Table 6.2	Odd ratios by factor . . . . .	122
Table 6.3	Stratified analysis by studied factors . . . . .	123

Table 6.4	Multiple logistic regression model . . . . .	124
Table 6.5	Odds ratios from multiple logistic regression . . . . .	124
Table 7.1	AV status over the 4 months . . . . .	139
Table 7.2	AV diversity over the 4 months . . . . .	141
Table 7.3	AV state changes over the 4 months . . . . .	142
Table 7.4	Overall AV stability . . . . .	143
Table 7.5	Descriptive statistics . . . . .	144
Table 7.6	Pearson correlation coefficients between infection rate and indicators (N=126 countries) . . . . .	145
Table 7.7	Pearson correlation coefficient matrix between indicators (N=126 coun- tries) . . . . .	145
Table 7.8	Multiple general linear regression (N=126 countries) . . . . .	146
Table 7.9	Pearson correlation coefficients between infection rates and indicators by protection status (N=126) . . . . .	147
Table 7.10	Multiple general linear regression by protection status (N=126 countries)	147
Table 8.1	Frequency of disease by cohort . . . . .	156
Table 8.2	Descriptive statistics by group . . . . .	159
Table 8.3	Frequency of malware infection by group . . . . .	159
Table 8.4	Estimates of overall AVE at 95% . . . . .	161
Table 8.5	AV protection status by user factor . . . . .	163
Table 8.6	User factors by HDI . . . . .	164
Table 8.7	Estimates of AVE at 95% . . . . .	165
Table 8.8	Estimates of AVE at 95% for different malware types . . . . .	166
Table 8.9	Estimates of AVE at 95% by gender . . . . .	167
Table C.1	Global multiple general linear regression results (N=50 countries) . .	202
Table D.1	Distribution of Windows versions by socio-economic status . . . . .	203
Table F.1	Odds ratios for gender by malware type . . . . .	205
Table F.2	Odds ratios for age by malware type . . . . .	206
Table G.1	Multiple logistic regression for adware . . . . .	207
Table G.2	Multiple logistic regression for virus . . . . .	207
Table G.3	Multiple logistic regression for cracks . . . . .	208
Table G.4	Multiple logistic regression for hack . . . . .	208
Table G.5	Multiple logistic regression for exploit . . . . .	209
Table G.6	Multiple logistic regression for rogue . . . . .	209
Table G.7	Multiple logistic regression for infostealer . . . . .	210
Table G.8	Multiple logistic regression for ransomware . . . . .	210

Table	G.9	Multiple logistic regression for bot . . . . .	211
Table	G.10	Multiple logistic regression for rootkit . . . . .	211
Table	H.1	Estimates of AVE at 95% by age group . . . . .	212
Table	H.2	Estimates of AVE at 95% by region . . . . .	213
Table	H.3	Estimates of AVE at 95% by HDI category . . . . .	213



## LISTE DES FIGURES

Figure 2.1	Modèle élargi des déterminants sociaux de la santé (Dahlgren et Whitehead, 1991) . . . . .	22
Figure 2.2	Modèle de prévention en santé publique . . . . .	24
Figure 3.1	Déterminants des attaques par logiciels malveillants . . . . .	37
Figure 4.1	Global map of malware infection rates . . . . .	59
Figure 4.2	Pareto charts by socio-economic status . . . . .	62
Figure 4.3	Box plot of infection rates by socio-economic status . . . . .	63
Figure 5.1	Frequency histogram of unique detections . . . . .	89
Figure 5.2	Unique malware detections per month . . . . .	90
Figure 5.3	Malware detections by type . . . . .	90
Figure 7.1	AV activity over the 4 months . . . . .	140
Figure 7.2	Overall AV vendor changes . . . . .	142
Figure 7.3	Infection rates over the 4 months . . . . .	143
Figure 7.4	Infection rates for unprotected function of the protection coverage . .	148
Figure 8.1	Cohort study design . . . . .	157
Figure 8.2	Comparative cohort study design . . . . .	165
Figure B.1	Residual analysis . . . . .	200
Figure B.2	Residual analysis without China . . . . .	200

## LISTE DES SIGLES ET ABRÉVIATIONS

AMTSO	Anti-Malware Testing Standards Organization
CCC	Convention sur la Cybercriminalité
CGI	Cyber Green Initiative
DSL	Digital subscriber line
FAI	Fournisseur d'accès Internet
GCI	Global Cybersecurity Index
HDI	Human Development Index
IP	Internet Protocol
JPCERT	Japan Computer Emergency Response Team
LAP	London Action Plan
MSRT	Malicious Software Removal Tool
PIB	Produit intérieur brut
PNUD	Programme des Nations Unies pour le développement
PPA	Parité de pouvoir d'achat
SIDA	Syndrome d'immunodéficience acquise
RAT	Routine activity theory
RTTL	Real Time Threat List
TIC	Technologies de l'information et de la communication
UIT	Union internationale des télécommunications
WINE	Worldwide Intelligence Network Environment

## LISTE DES ANNEXES

Annexe A	DESCRIPTION OF COUNTRY-LEVEL FACTORS . . . . .	199
Annexe B	RESIDUAL ANALYSIS . . . . .	200
Annexe C	MULTIPLE GENERAL LINEAR REGRESSION RESULTS . . . . .	202
Annexe D	WINDOWS VERSIONS STATISTICS . . . . .	203
Annexe E	DEFINITIONS BY MALWARE TYPE . . . . .	204
Annexe F	ODDS RATIOS BY MALWARE TYPE . . . . .	205
Annexe G	MULTIPLE LOGISTIC REGRESSION BY MALWARE TYPE . . . . .	207
Annexe H	ESTIMATES OF AVE AT 95% . . . . .	212

## CHAPITRE 1 INTRODUCTION

Nous présentons dans ce premier chapitre le contexte dans lequel s'inscrit notre projet de recherche. Plus spécifiquement, nous exposons les éléments de la problématique étudiée ainsi que les principales motivations sous-jacentes à la présente thèse. Finalement, nous énonçons les principaux objectifs de recherche qui ont guidé la démarche de notre travail, et nos principales contributions en termes d'avancement des connaissances et de publications scientifiques.

### 1.1 Éléments de la problématique

Avec plus de trois milliards d'utilisateurs connectés à Internet, les fondations de notre société moderne reposent de plus en plus sur le numérique. Le développement des infrastructures numériques a contribué, notamment, à multiplier les opportunités de communication, de collaboration, et de commerce. Le trafic global Internet annuel a franchi la barre des zettabytes (ZB) en 2016, et devrait atteindre 3.3 ZB par année d'ici l'an 2021 (Cisco Systems, 2017b). De 2.3 appareils connectés par habitant en 2016, ce nombre devrait passer à 3.5 d'ici 2021 (Cisco Systems, 2017b). Quant au nombre d'utilisateurs Internet, il est estimé atteindre 4.6 milliards d'ici 2021, soit plus de 50% de la population globale (Cisco Systems, 2017a).

Cette augmentation du nombre de personnes, d'appareils et de données a notamment contribué à augmenter la motivation et l'opportunité des attaques informatiques. En quelques années, le nombre de nouveaux fichiers malveillants observés par la compagnie Panda Security serait passer de 230,000 fichiers par jour en 2015 à 285,000 en 2017 (PandaLabs, 2017). De son côté, la compagnie Kaspersky rapporte avoir observé 323,000 nouveaux fichiers malveillants par jour en 2016, représentant une augmentation de 13,000 en comparaison avec 2015 (Kaspersky Lab, 2016). En outre, le nombre de logiciels de rançon, un type de logiciels malveillants qui bloque l'accès à des ressources informatiques dans le but d'extorquer de l'argent à son utilisateur, aurait augmenté d'un facteur de 30 entre 2015 et 2016 (Proofpoint, 2016). À eux seuls, les dommages causés par les logiciels de rançon auraient excédé les 5 milliards de dollars américains uniquement en 2017 (Morgan, Steve, 2017). Pour sa part, le nombre total de vols de données reportés publiquement serait passé de 136 en 2005, à plus de 7,885 en 2018 (Privacy Rights ClearingHouse, 2018). Selon la compagnie Juniper Research, le coût cumulatif des vols de données devrait atteindre un total de 8 trillions de dollars américains pour la période de 2017 à 2022 (Moar, James, 2017). Quant au coût annuel global des attaques informatiques, il est estimé atteindre les 6 trillions de dollars américains d'ici 2021 (Morgan, Steve, 2017), en faisant le plus grand transfert de richesse économique de l'histoire.

En réponse, la communauté en sécurité des systèmes d'information a fait des progrès notables dans le développement et l'implémentation de solutions techniques, sociales, et légales pour réduire le risque d'attaques informatiques. Notamment, le projet *No More Ransom*, créé en 2016 par un regroupement d'agences de polices et de sociétés spécialisées en sécurité informatique, vise à combattre les entreprises cybercriminelles ayant des connections avec les logiciels de rançon (No More Ransom Project, 2018). Un autre exemple notable concerne l'avènement des antivirus dits *nouvelle génération* (*next-gen* en anglais). Alors que les antivirus traditionnels reposent principalement sur des signatures et des heuristiques, la nouvelle génération se distingue, entre autres, par son utilisation de nouvelles techniques de détection basées sur l'intelligence artificielle. Cependant, la majorité des efforts en matière de prévention et de protection sont dédiés au développement de nouvelles solutions et peu d'attention est accordée à leur évaluation et leur adoption. Considérant la croissance de l'Internet et des menaces informatiques, il devient essentiel pour les gouvernements, les entreprises, et les individus de définir des modèles et des moyens de coopération pour identifier et évaluer les interventions visant à réduire le risque que posent les attaques informatiques.

Ce manque d'approche globale et de coordination a amené plusieurs chercheurs à explorer l'analogie entre la santé et la sécurité des systèmes d'information. Notamment, plusieurs travaux de recherche ont proposé de s'inspirer de l'approche et des méthodes utilisées en *santé publique* (Rice *et al.*, 2010; Rowe *et al.*, 2012b; Sullivan *et al.*, 2012; Rowe *et al.*, 2013). Cette approche, qui repose sur un ensemble de dimensions (administratives, sociales, politiques, et économiques), s'occupe de tous les aspects et moyens mis en place pour préserver et promouvoir la santé. Par exemple, Sullivan *et al.* (2012) ont proposé un cadre conceptuel visant à protéger la *santé* d'Internet des attaques par logiciels malveillants et d'autres menaces informatiques. Rice *et al.* (2010) ont quant à eux développé une stratégie globale visant à préserver la *santé* du cyberspace en réduisant l'occurrence et l'impact des infections par logiciels malveillants.

Similairement, nous croyons que le domaine de la sécurité des systèmes d'information peut bénéficier des leçons apprises et des méthodes utilisées dans le domaine de la santé. Par exemple, nous savons que la santé est affectée par plusieurs niveaux de facteurs (global, environnemental, individuel, etc.) qui sont reliés les uns aux autres. Or, ce sont les interventions visant à changer le contexte global et l'environnement qui se sont révélées être les plus efficaces. Contrairement aux interventions visant directement les individus, telles que l'éducation et les interventions cliniques, les interventions à plus haut niveau permettent d'atteindre une plus grande partie de la population tout en minimisant les efforts individuels. Bien qu'il soit nécessaire de développer des interventions à tous les niveaux, l'expérience a démontré qu'une intervention a priori efficace peut ne pas avoir le succès escompté si le contexte spécifique

de la population et des individus n'a pas été pris en compte lors de son implémentation. À l'instar de la santé, l'adoption d'une perspective multi-niveaux axée sur la santé publique peut fournir un cadre global pour protéger, maintenir, et améliorer la sécurité des systèmes d'information. En particulier, une telle approche permettrait d'encadrer le développement d'interventions visant à prévenir et/ou réduire le risque d'attaques informatiques.

## 1.2 Motivations

Tel que mentionné dans la section précédente, une approche de prévention basée sur le modèle de la santé publique permettrait de i) identifier les facteurs qui affectent la sécurité des systèmes d'information et en comprendre les causes sous-jacentes, ii) développer et évaluer des stratégies visant à améliorer la sécurité des systèmes d'information, et iii) implémenter et disséminer auprès de la population les stratégies développées.

**Identification de facteurs de risque** L'approche de la santé publique peut permettre d'identifier et de mieux comprendre quels sont les causes et les corrélats qui déterminent la sécurité des systèmes d'information. En particulier, une telle approche permettrait de i) faire ressortir la dimension multi-niveaux et les rôles interactifs des facteurs de risque, ii) identifier les populations à risque et comprendre comment ces dernières sont différemment affectées, et iii) orienter les efforts de recherche visant à identifier les facteurs qui affectent la sécurité des systèmes d'information.

**Développement de stratégies** Une meilleure compréhension des facteurs de risque et des facteurs de protection reliés à la sécurité des systèmes d'information peut ainsi permettre de développer des stratégies efficaces basées sur des faits et des données probantes. Qui plus est, une approche basée sur la santé publique permettrait de i) développer des interventions dites *écologiques* qui prennent en compte le contexte des individus, des communautés, et de la population, et ii) considérer la toile causale, soit le lien entre les composantes visées par l'intervention et l'atteinte des résultats désirés.

**Évaluation de stratégies** À l'instar de la santé, l'évaluation des stratégies en sécurité des systèmes d'information devrait être continue. En d'autres mots, chaque stratégie devrait être rigoureusement évaluée avant, pendant, et après son implémentation. Plus particulièrement, l'adoption d'un cadre d'évaluation inspiré du modèle de la santé publique permettrait de : i) développer des méthodologies permettant de prédire l'impact des interventions avant de les déployer, ii) mettre en évidence la relation entre les investissements et les résultats obtenus

en termes d'efficacité, d'utilité et de bénéfices, et iii) guider les efforts de priorisation en sélectionnant les stratégies les plus efficaces dans un contexte donné.

**Implémentation et adoption de stratégies** Les stratégies prouvées comme étant efficaces doivent par la suite être implémentées afin d'en garantir les résultats escomptés. À cet effet, l'approche de la santé publique appliquée à la sécurité des systèmes d'information peut permettre de : i) identifier les méthodes de dissémination les plus efficaces dans un contexte donné, ii) guider les efforts de dissémination afin de cibler, par exemple, les populations les plus à risque, et iii) comprendre les facteurs qui empêchent et/ou limitent l'adoption des stratégies afin d'y remédier.

### 1.3 Objectifs de recherche

Nous proposons de nous inspirer des méthodes en santé publique pour développer un modèle de prévention applicable à la sécurité des systèmes d'information. En particulier, notre travail se concentre sur un type de menaces informatiques : les attaques par logiciels malveillants. Ce choix est principalement motivé par i) l'étendue du problème, ii) la magnitude de l'impact des attaques, et iii) la croissance rapide et l'évolution des logiciels malveillants au cours des dernières années. Dans ce contexte, l'objectif général de la présente thèse est de :

- *Développer et appliquer un modèle basé sur l'approche de la santé publique pour la prévention des attaques par logiciels malveillants.*

Quant à l'application du modèle de prévention, deux objectifs spécifiques guident notre démarche :

- *Identifier les causes et les corrélats reliés aux attaques par logiciels malveillants. Plus spécifiquement, déterminer quels sont les facteurs de risque et les facteurs de protection qui influencent le risque d'attaques par logiciels malveillants.*
- *Évaluer l'efficacité réelle d'une intervention visant à prévenir et/ou réduire l'occurrence des attaques par logiciels malveillants. Plus spécifiquement, évaluer l'efficacité d'une intervention technique largement utilisée, soit les solutions antivirus.*

## 1.4 Contributions

La présente thèse propose de s’inspirer de l’approche axée sur la santé publique pour développer un modèle de prévention applicable à la sécurité des systèmes d’information. À cet effet, nos principales contributions au domaine de la sécurité des systèmes d’information peuvent être résumées comme suit :

**Modèle de prévention** Développement d’un modèle axé sur l’approche de la santé publique pour la prévention des attaques par logiciels malveillants. Ce modèle est décrit au Chapitre 3.

**Facteurs socio-environnementaux** Réalisation d’une étude écologique multi-pays visant à identifier les différents facteurs socio-environnementaux reliés aux infections par logiciels malveillants. Les principales contributions de cette étude sont : i) l’identification et la quantification de l’effet relatif de facteurs socio-environnementaux sur le taux national d’infections par logiciels malveillants, ii) la mise en évidence de la variation de l’effet de ces facteurs en fonction de différents statuts socio-économiques, et iii) l’identification de facteurs de risque modifiables qui peuvent être influencés par des politiques nationales en matière de sécurité des systèmes d’information. Ces résultats sont présentés au Chapitre 4, et ont été publiés en 2016 à la conférence *Workshop on the Economics of Information Security (WEIS)* (Lalonde Lévesque *et al.*, 2016).

**Facteurs comportementaux** Analyse de facteurs comportementaux reliés au risque d’attaques par logiciels malveillants. Cette analyse fait suite à l’étude utilisateurs réalisée par Lalonde Lévesque *et al.* (2013), qui visait à étudier l’interaction entre les usagers, les solutions antivirus, et les logiciels malveillants. L’analyse réalisée dans le cadre de cette thèse a permis, quant à elle, d’identifier des facteurs de risque reliés au comportement usager. Cette analyse a fait l’objet d’un article qui a été publié en 2018 par la revue *ACM Transactions on Privacy and Security* (Lalonde Lévesque *et al.*, 2018). Les détails de cette étude sont présentés au Chapitre 5.

**Facteurs démographiques** Réalisation d’une étude cas-témoins visant à identifier les populations d’usagers les plus à risque, et étudier comment l’effet de facteurs démographiques, soit l’âge et le genre, varie pour différents types de logiciels malveillants. Cette étude, soit la première dédiée à analyser la relation entre l’âge, le genre, et l’exposition aux logiciels malveillants, a permis de : i) identifier l’âge et le genre comme étant des corrélats significatifs



reliés au risque d’attaques par logiciels malveillants, ii) identifier des populations usagers à risque plus élevé, et iii) mettre en évidence comment l’effet de l’âge et du genre varient en direction et en magnitude selon le type de logiciels malveillants. Les résultats de cette étude ont été publiés en 2017 à la *British Human Computer Interaction Conference (BHCI)* (Lalonde Lévesque *et al.*, 2017). Ils sont décrits au Chapitre 6.

**Évaluation globale de solutions antivirus** Réalisation d’une étude permettant d’évaluer la *santé* de l’écosystème des logiciels antivirus, et de mesurer la performance réelle agrégée de ce dernier. Cette analyse, soit une première en son genre, a permis de : i) développer et identifier des indicateurs reliés à la santé de l’écosystème des logiciels antivirus, et ii) identifier des secteurs sujets à amélioration. Ces résultats sont décrits au Chapitre 7. Ils ont été publiés en 2015 à la conférence *International conference on Malicious and Unwanted Software : The Americas (MALWARE)* (Lalonde Lévesque *et al.*, 2015).

**Évaluation comparative de solutions antivirus** Réalisation d’une étude cohorte visant à évaluer l’efficacité réelle des logiciels antivirus, et de mesurer comment cette dernière est affectée par différents facteurs externes, tels que le type de logiciel malveillant, le profil démographique de l’utilisateur, et l’environnement socio-économique. À cet effet, nos principales contributions sont : i) le développement d’une méthodologie permettant d’évaluer en conditions réelles la performance de solutions antivirus, ii) la réalisation d’une première évaluation comparative de solutions antivirus en conditions réelles, et iii) la mise en évidence de l’impact de différents facteurs externes sur la performance réelle des solutions antivirus. Ces résultats ont été publiés en 2016 à la conférence *International Virus Bulletin Conference* (Lalonde Lévesque *et al.*, 2016a). Ils sont décrits au Chapitre 8.

## 1.5 Publications

La section qui suit présente la liste des articles scientifiques issus de notre travail de recherche.

### Identification de facteurs reliés aux attaques par logiciels malveillants

- **F. Lalonde Lévesque**, Sonia Chiasson, A. Somayaji, J.M. Fernandez, “Technological and human factors of malware attacks : a computer security clinical trial approach”, *ACM Transactions on Privacy and Security*, Juin, 2018. (**Chapitre 5**)
- Jude Jacob Nsiempba, **F. Lalonde Lévesque**, Nathalie De Marcellis-Warin, J.M. Fernandez, “Short paper : An empirical analysis of risk aversion in malware infec-

tions”, *International Conference on Risks and Security of Internet and Systems*, Dinard, France, 19-21 Septembre 2017.

- **F. Lalonde Lévesque**, J.M. Fernandez, D. Batchelder, “Age and gender as independent risk factors for malware victimisation”, *British Human Computer Interaction Conference*, Sunderland, Royaume-Uni, 3-6 Juillet 2017. (**Chapitre 6**)
- A. Arrott, **F. Lalonde Lévesque**, D. Batchelder, J.M. Fernandez, “Citizen Cybersecurity Health Metrics for Windows Computers”, *Central and Eastern European eDem and eGov Days*, Budapest, Hongrie, 12-13 Mai, 2016.
- **F. Lalonde Lévesque**, J.M. Fernandez, A. Somayaji, D. Batchelder, “National-level risk assessment : A multi-country study of malware infections”, *Workshop on The Economics of Information Security*, Berkeley, États-Unis, 13-14 Juin, 2016. (**Chapitre 4**)

## Évaluation de solutions antivirus à prévenir les attaques par logiciel malveillants

- **F. Lalonde Lévesque**, J.M. Fernandez, D. Batchelder, G. Young, “Are they real? Real-life comparative tests of anti-virus products”, *Virus Bulletin International Conference*, Denver, États-Unis, 5-7 Octobre, 2016. (**Chapitre 8**)
- **F. Lalonde Lévesque**, A. Somayaji, D. Batchelder, J.M. Fernandez, “Measuring the health of antivirus ecosystem”, *IEEE International Conference on Malicious and Unwanted Software*, Fajardo, États-Unis, 20-22 Octobre, 2015. (**Chapitre 7**)

## 1.6 Plan de la thèse

La présente thèse est organisée en quatre parties distinctes. La première partie vise à introduire le lecteur aux différents concepts et travaux antérieurs nécessaires à la bonne compréhension du travail de recherche. Le Chapitre 2 présente les principaux concepts associés aux logiciels malveillants, aux solutions antivirus, et à la santé publique. Le Chapitre 3 présente le développement d’un modèle de prévention pour la sécurité des systèmes d’information. En particulier, nous élaborons les détails de son application au contexte spécifique des attaques par logiciels malveillants.

La deuxième et la troisième partie de la thèse, quant à elles, portent sur les travaux réalisés dans le cadre de nos deux objectifs spécifiques. La deuxième partie couvre le premier objectif, soit identifier les causes et les corrélats reliés aux attaques par logiciels malveillants. Le Chapitre 4 se concentre sur les facteurs socio-environnementaux, le Chapitre 5 sur les facteurs comportementaux, et le Chapitre 6 sur les facteurs démographiques. Pour sa part, la troisième partie couvre le second objectif spécifique, soit évaluer l’efficacité réelle des solutions antivirus à prévenir et/ou réduire l’occurrence des attaques par logiciels malveillants.

Le Chapitre 7 présente les résultats d’une évaluation agrégée des antivirus, et le Chapitre 8 présente les travaux reliés à une évaluation comparative des solutions antivirus.

La quatrième partie vient conclure la présente thèse. Le Chapitre 9 vient mettre en contexte et discuter les différents résultats obtenus lors de la réalisation des travaux de recherche. Finalement, le Chapitre 10 présente une synthèse des travaux réalisés, et des travaux futur.

## CHAPITRE 2 SÉCURITÉ ET SANTÉ PUBLIQUE

Le chapitre qui suit se veut une introduction aux différents concepts de base relatifs aux logiciels malveillants, aux logiciels antivirus, ainsi qu'à l'approche de la santé publique. Finalement, nous présentons les travaux de recherche existants qui ont exploré l'analogie entre la santé publique et la sécurité des systèmes d'information.

### 2.1 Logiciels malveillants

Un logiciel, ou programme informatique, est considéré malveillant s'il vise à détruire, endommager ou détourner l'utilisation légitime d'un système informatique et ce, sans le consentement de l'utilisateur. À titre d'exemples, nous entendons par logiciel malveillant, les virus, les vers, les chevaux de Troie et autres menaces. Dans la section qui suit, nous présentons une introduction aux principaux concepts relatifs aux logiciels malveillants, soit leur classification, leurs moyens de propagation, ainsi que les principales motivations associées aux attaques par logiciels malveillants.

#### 2.1.1 Classification

Les logiciels malveillants sont généralement classifiés en fonction de leurs caractéristiques et de leur comportement. Nous présentons ci-dessous une description de certains types de logiciels malveillants parmi les plus communs.

**Virus informatique** Les virus informatiques tirent leur nom d'une analogie avec les virus biologiques. Tout comme ces derniers, ils se reproduisent en infectant un programme légitime ou un document qui agit à titre d'hôte. Ce type de logiciel malveillant peut être classifié selon le type d'objet infecté. Par exemple, un virus infecteur de fichier, est un type de virus qui infecte les programmes présents sur un système informatique. Un autre type de virus, le virus de secteur d'amorçage, s'attache au secteur de démarrage d'un support amovible, tel qu'un cédérom, une clef USB, ou un disque dur.

**Ver informatique** Les vers informatiques, *computer worms* en anglais, sont des logiciels malveillants autonomes. À l'opposé des virus informatiques, ils ne nécessitent pas la présence d'un hôte pour se reproduire. Différents types de vers informatiques sont définis en fonction de leur moyen de propagation. Par exemple, certains vers informatiques se propagent par envoi

automatique de courrier électronique, alors que d'autres vont analyser un réseau informatique afin d'identifier les systèmes vulnérables.

**Cheval de Troie** Le terme cheval de Troie, *trojan horse* en anglais, provient d'une référence à l'épopée d'Illiade de Homère. Ce type de programme informatique d'apparence légitime exécute des actions malveillantes sur un système à l'insu de son utilisateur. Contrairement aux virus et aux vers informatiques, les chevaux de Troie ne peuvent se reproduire ou se propager sans assistance.

**Logiciel publicitaire** Les logiciels publicitaires, *adware* en anglais, sont des programmes informatiques qui affichent de la publicité lors de leur installation et/ou lors de leur utilisation. Légitimes à la base, certains logiciels publicitaires sont considérés comme des menaces informatiques puisqu'ils présentent des publicités non désirées par l'utilisateur, ou qu'ils collectent des informations sur ce dernier. Dans un tel cas, le programme informatique sera qualifié de logiciel espion.

**Logiciel espion** Tel que mentionné précédemment, les logiciels espions, *spyware* en anglais, sont des programmes qui visent à collecter sans autorisation des informations sur l'utilisation d'un système.

**Logiciel de sécurité falsifié** Comme son nom l'indique, un logiciel de sécurité falsifié, *rogueware* en anglais, est un faux logiciel de sécurité. Ce type de logiciel malveillant tente de convaincre l'utilisateur que son système est infecté afin de l'inviter à acheter un (faux) logiciel de sécurité. Dans certains cas, le téléchargement du logiciel de sécurité est utilisé comme vecteur d'infection afin d'installer un autre logiciel malveillant.

**Logiciel de rançon** Un logiciel de rançon, *ransomware* en anglais, est un programme informatique qui bloque l'accès à des ressources informatiques dans le but d'extorquer de l'argent à son utilisateur. Les logiciels de rançon sont classifiés en deux catégories, selon qu'ils utilisent, ou pas, le chiffrement. Dans le premier cas, le programme va recourir au chiffrement et exiger une rançon en échange du déchiffrement. Dans le second cas, le programme va restreindre toute interaction avec le système, et forcer l'utilisateur à payer une rançon pour récupérer l'accès aux ressources prises en otage.

**Bot** Le nom de ce type de logiciel est dérivé du mot *robot*. Il désigne un programme installé sur un système afin de réaliser automatiquement des actions spécifiques. Dans un contexte

malveillant, ces logiciels peuvent être utilisés afin de réaliser des attaques par déni de service, envoyer du spam massivement, voler des informations, ou encore miner de la crypto-monnaie.

### 2.1.2 Moyens de propagation

On entend par moyens de propagation les méthodes électroniques de transmission utilisées par les logiciels malveillants pour infecter un système. Par exemple, certains logiciels malveillants vont exploiter des vulnérabilités logicielles afin de se propager et infecter les systèmes. D'autres vont reposer sur l'interaction de l'utilisateur, ou recourir à une combinaison des deux techniques. Compte tenu des nombreuses méthodes de propagation utilisées par les logiciels malveillants, nous présentons dans la section qui suit les méthodes les plus communes.

**Web** L'utilisation de sites Web afin de distribuer des logiciels malveillants est une des méthodes de propagation les plus fréquentes. Pour ce faire, l'attaquant peut soit compromettre un site Web légitime et y héberger des logiciels malveillants, ou configurer un site Web spécialement enregistré à des fins malveillantes. Dans certains cas, le simple fait de visiter un site Web infecté suffit à compromettre le système. Dans d'autres cas, l'utilisateur peut être amené à cliquer sur une fenêtre pop-up qui enclenchera le téléchargement d'un logiciel malveillant. Cette technique, soit le téléchargement furtif, *drive-by download* en anglais, est basé sur l'exploitation de vulnérabilités logicielles et/ou sur l'ingénierie sociale. Une autre technique, soit le *malvertising* en anglais, consiste à injecter des logiciels malveillants dans des publicités en ligne. Contrairement aux scénarios d'attaques précédents, l'attaquant n'a pas besoin de compromettre ou créer un site Web.

**Courrier électronique** Les logiciels malveillants peuvent être envoyés par l'intermédiaire de courriers électroniques. Dans un cas, le programme malveillant peut être directement intégré au courrier électronique et infecter l'utilisateur à son insu en exploitant des vulnérabilités logicielles au niveau du client courriel, par exemple. Il s'agit alors d'un téléchargement furtif. Dans un autre cas, le logiciel malveillant peut être attaché au courrier électronique sous la forme d'un fichier (binaire, vidéo, audio, image, etc.). L'attaquant peut alors combiner différentes techniques d'ingénierie sociale afin d'amener l'utilisateur à télécharger/visionner le fichier joint.

**Messagerie instantannée** Similairement aux courriers électroniques, les services de messagerie instantannée peuvent être exploités par les logiciels malveillants comme moyen de propagation. Le message envoyé peut soit comprendre un fichier malveillant, ou simplement

contenir un lien vers un site Web hébergeant des logiciels malveillants.

**Réseaux sociaux** Le recours aux réseaux sociaux, tels que Facebook, Twitter, ou LinkedIn, peut permettre aux logiciels malveillants de se propager dans un contexte où les utilisateurs sont souvent moins méfiants. Par exemple, du code malveillant peut être intégré dans une image et exécuté lorsque l'utilisateur télécharge cette dernière. Cette technique, baptisée *ImageGate* par la firme de sécurité Check Point, a notamment été utilisée par le logiciel de rançon Locky afin d'infecter des milliers d'utilisateurs (Zaikin et Barda, 2016). D'autres logiciels malveillants vont utiliser différentes techniques d'ingénierie sociale afin d'amener les utilisateurs à cliquer sur un lien pointant vers un site Web malveillant.

**Réseaux locaux sans-fil** Utilisés à la base pour créer une liaison entre plusieurs ordinateurs, les réseaux locaux sans fil peuvent être exploités par les logiciels malveillants. Par exemple, un logiciel malveillant peut analyser un réseau afin d'identifier et d'infecter les systèmes connectés identifiés comme étant vulnérables.

**Réseaux pair à pair/partage de fichiers** Les réseaux pair à pair, *peer-to-peer* en anglais, permettent à plusieurs ordinateurs de partager des données via un réseau et ce, sans transiter par un serveur central. Il est possible pour un attaquant de dissimuler un fichier malveillant sous la forme d'une vidéo populaire, ou encore d'une application légitime. De là, il ne reste qu'à convaincre l'utilisateur de télécharger et d'exécuter le fichier. Un logiciel malveillant peut aussi se propager en se copiant dans un dossier de fichiers partagés. Les utilisateurs qui exécutent le fichier partagé seront alors infectés.

**Appareil et média amovible** Un logiciel malveillant installé sur un appareil ou média amovible (clef USB, disque dur externe, carte mémoire, etc.) peut infecter et/ou se propager en s'exécutant automatiquement dès qu'il est connecté à un système.

### 2.1.3 Acteurs malveillants et motivations

**Attaquants organisés** Les attaquants dits organisés incluent les organisations terroristes, les hacktivistes, les gouvernements, et les organisations criminelles. Les organisations terroristes visent, entre autres, à faire une déclaration politique ou infliger à leur cible des dommages psychologiques et/ou physiques dans le but d'obtenir un gain politique ou d'inciter la peur. Similairement, les hacktivistes cherchent souvent à faire des déclarations politiques. Cependant, contrairement aux terroristes, leur but premier est de sensibiliser, et non d'encou-

rager le changement par la peur. Quant aux gouvernements, leurs principaux motifs visent le vol d'information ou le sabotage. Ces attaques sont financées et réalisées par des pirates hautement qualifiés. Pour ce qui est des organisations criminelles, elles sont souvent constituées de criminels professionnels. Ces derniers sont généralement attirés par le contrôle, le pouvoir, et l'argent.

**Pirates informatique** Les pirates informatiques peuvent être considérés bénins ou malveillants. Dans le premier cas, le pirate peut être qualifié de “chapeau blanc”, *white hat* en anglais, et se livrer à des activités de piratage dans le but de découvrir des vulnérabilités dans les systèmes informatiques. Dans le second cas, le pirate informatique sera qualifié de “chapeau noir”, soit *black hat* en anglais. Ce type d'attaquant prend part à des activités de piratage à des fins illégales. Ils peuvent soit agir sur une base individuelle, ou être engagés par des organisations criminelles ou encore des gouvernements. Leurs objectifs peuvent varier entre l'espionnage, le vol d'information, l'extorsion, ou encore le vandalisme.

**Amateurs** Ce type d'attaquant, appelé *script kiddies* ou *noobs* en anglais, est souvent qualifié de pirate informatique moins expérimenté ; ils utilisent principalement des outils et méthodes disponibles sur Internet. Malgré la simplicité de leurs attaques, ces dernières peuvent tout de même causer énormément de dommages. Leurs principales motivations peuvent inclure la curiosité, le défi personnel, ou la démonstration de leurs compétences.

## Motivations

À partir des différents acteurs malveillants et de leurs objectifs, nous avons classifié les motivations en trois catégories, soit les motivations économiques, politiques, et socio-culturelles.

**Économiques** Les motifs d'ordre économique sont à l'origine d'une grande majorité des attaques par logiciels malveillants. En termes d'objectifs financiers, nous retrouvons le vol de propriété intellectuelle ou d'informations bancaires, et l'espionnage industriel. Notamment, certains Trojans dits *bancaires*, tels que Zeus, Neverquest, Gozi et Dridex, se spécialisent dans le vol d'informations bancaires. Par exemple, le trojan bancaire Dridex, qui cible 315 institutions bancaires (Wueest, 2015), a causé des dommages estimés à plus de 40 millions de dollars américains uniquement en 2015 (Slepogin, 2017). Alors que certaines attaques visent à voler les informations bancaires des clients de certaines banques, d'autres ciblent directement les banques et les entreprises. Un autre objectif, l'extorsion, se retrouve à la base des attaques par logiciels de rançon. Selon un rapport de la compagnie Symantec,



les montants demandés en 2017 se situent en moyenne à 500\$ par rançon (O'Brien, 2017). Quant au nombre de victimes qui payent les montants demandés, ce dernier est estimé à 40%, totalisant plus de 209 million de dollars américains extorqués uniquement durant le premier quart de l'année 2016 (O'Brien, 2016). Un autre objectif financier, particulièrement à la hausse depuis 2017, consiste à infecter des systèmes afin d'utiliser leurs ressources pour miner de la crypto-monnaie, telles que Bitcoin et Monero. À cet effet, le groupe Anti-Phishing Working Group rapporte que près de 1.2 milliard de dollars américain auraient été volés en crypto-monnaie et ce, entre janvier 2017 et mai 2018 (Chavez-Dreyfus, 2018).

**Politiques** Les attaques de nature politique peuvent viser à détruire ou perturber une cible. Par exemple, le logiciel malveillant *Industroyer* a ciblé en 2016 le réseau électrique d'Ukraine, privant ainsi d'électricité pendant une heure Kiev, la capitale (Cherepanov, Anton, 2017). D'autres attaques ont comme objectif d'espionner, faire des déclarations politiques, des protestations ou encore des actions de représailles. Notamment, le groupe *Sednit*, connu aussi sous le nom *APT28*, *Fancy Bear*, *Sofacy*, et *Pawn Storm*, est un groupe d'attaquants organisés qui opère depuis 2004 dans le but de voler des informations confidentielles (Eset, 2016). Ce groupe a, entre autres, ciblé en 2017 les élections présidentielles en France en tentant d'installer un logiciel malveillant sur le site Web de la campagne électorale d'Emmanuel Macron.

**Socio-culturelles** Les attaques basées sur des motifs socio-culturelles ont souvent des objectifs philosophiques, politiques, ou humanitaires. Le groupe les Anonymes, *Anonymous* en anglais, est notamment connu pour ses nombreuses attaques à caractère hacktiviste. Par exemple, le groupe a piraté en 2013 le site Web du gouvernement nigérien suite à l'adoption d'une loi visant à punir les relations homosexuelles (Ogala, 2013). En 2017, c'est le site Web du groupe terroriste ISIS qui a été victime d'une attaque visant à y installer un logiciel malveillant (Hassan, 2017). Les motivations socio-culturelles peuvent aussi comprendre la curiosité, l'amusement, la recherche de visibilité, ou encore la gratification de l'ego.

#### 2.1.4 Tendances et prévalence des logiciels malveillants

##### Évolution des attaques

Le paysage des logiciels malveillants des dernières années (2012-2017) a principalement été dominé par les chevaux de Troie. En 2012, quatre nouveaux fichiers malveillants détectés sur cinq par la compagnie Panda Labs appartenaient à la catégorie des chevaux de Troie (Panda Labs, 2012). Similairement, la corporation Microsoft rapporte que les chevaux de Troie, sui-

vis par les vers informatiques, sont les types de logiciels malveillants les plus prévalents depuis 2012 (Microsoft Corporation, 2014a, 2013a, 2014d; Corporation, 2017). Cependant, ces tendances semblent changer selon la compagnie, les données observées, et la période. Par exemple, pour AV-Test, ce sont les virus traditionnels, les vers informatiques, et les trojans qui se volent respectivement la vedette depuis 2012 (AV-TEST, 2016, 2017). L'année 2103 a pour sa part été marquée par une augmentation des logiciels de rançon, plus particulièrement des logiciels à chiffrement, tels que Locky et CryptoLocker. Ce dernier a notamment réussi à infecter plus de 500 000 systèmes entre septembre 2013 et mai 2014 et ce, par téléchargement furtif, et envoi de pièce jointe malveillante. En 2017, les logiciels de rançon adoptent une nouvelle technique; celle de s'auto-propager tel un ver informatique. Notamment, les logiciels WannaCry et Petya se sont démarqués de leurs prédécesseurs de par leur intention d'endommager les données et les systèmes, en plus de réclamer une rançon. 2017 a de plus été marquée par une diminution du nombre de kits d'exploitation de vulnérabilités; *Angler*, *Disdain* et *Terror* ont disparu, *Neutrino* est devenu privé, et *Sundown* a cessé d'offrir ses services (Trend Micro, 2017). Quant à la propagation des logiciels malveillants, les moyens de transmission les plus fréquents entre 2012 et 2017 sont, selon la corporation Microsoft, les réseaux pair à pair/le partage de fichiers, les sites Web, et le courrier électronique (Microsoft Corporation, 2013b,a, 2014d, 2015).

### **Prévalence et distribution des logiciels malveillants**

Selon un rapport de la corporation Microsoft, 10.3% des ordinateurs protégés par les solutions de protection Windows temps-réel ont expérimenté au moins une attaque par logiciel malveillant et ce, uniquement durant le mois de janvier 2017 (Corporation, 2017). Pour sa part, la compagnie de sécurité Kaspersky rapporte qu'en moyenne 20.5% de leurs clients ont été exposés à une attaque par logiciel malveillant durant le premier quart de l'année 2017 (Unucheck *et al.*, 2017). Au total, les produits de Kaspersky ont détecté et/ou bloqué des attaques par logiciels malveillants sur 29.4% de leurs clients durant l'année 2017 (Kaspersky Lab, 2017). En comparaison, ces figures semblent similaires aux taux d'attaques observés en 2012; la compagnie Panda Security a rapporté avoir bloqué des attaques par logiciels malveillants chez 31.63% de ses utilisateurs durant le deuxième quart de l'année (Panda Security Labs, 2012).

Quant à la distribution des attaques par logiciels malveillants, cette dernière révèle d'importantes variations d'ordre géographique. Par exemple, 37.67% des utilisateurs protégés par les solutions de Kaspersky en Algérie ont expérimenté une attaque malveillante au cours du premier quart de l'année 2017. À l'opposé, seulement 9.18% des utilisateurs au Japon

ont été exposés à une attaque malveillante au cours de la même période (Unuchek *et al.*, 2017). Un phénomène géographique similaire est observé au niveau des continents. L'Afrique indique parmi les taux d'attaques les plus élevés, alors que l'Amérique du Nord et l'Australie présentent les taux les moins élevés (Unuchek *et al.*, 2017; Corporation, 2017; Kleiner *et al.*, 2013). Ces patrons géographiques tendent cependant à varier selon le type de logiciels malveillants. Alors que l'Amérique du Nord présente historiquement les taux d'attaques les plus bas, l'inverse est observé, par exemple, dans le cadre des logiciels de rançon (Microsoft Secure, 2017).

Au niveau des systèmes d'opération, l'écosystème Windows présente historiquement les plus haut taux d'attaques ainsi que le plus grand nombre de logiciels malveillants. Selon un rapport de AV-TEST, 67.21% des fichiers malveillants détectés en 2017 visaient le système Windows. À l'opposé, seulement 0.07% des fichiers visaient macOS, et 0.02% le système Linux (AV-TEST, 2016). Cependant, depuis quelques années, ces chiffres sont à la baisse pour Windows, et inversement, à la hausse pour macOS. Notamment, le nombre de logiciels malveillants présents sur macOS est passé de 819 en 2015 à 3033 en 2016. Bien que ces chiffres soient bien en deça des 600 millions de logiciels malveillants observés sur Windows, ils indiquent tout de même une tendance vers la hausse pour le système développé par Apple (AV-TEST, 2017). Une analyse similaire au niveau de l'écosystème Windows révèle d'importantes variations entre les versions du système d'opération. Les versions les plus récentes du système de Microsoft présentent des taux d'attaques plus bas en comparaison avec les versions les plus anciennes. Par exemple, les taux d'infection de Windows 8.1 et 8 étaient estimés à 1.8% et 6.7% respectivement pour 2014, alors que le taux d'attaques de Windows Vista atteignait 10.4% (Microsoft Corporation, 2014d).

## 2.2 Logiciels antivirus

Les logiciels antivirus sont considérés comme étant la première ligne de défense contre les logiciels malveillants. Ces programmes informatiques ont entre autres pour fonctions d'identifier, d'arrêter et de supprimer les logiciels malveillants présents sur différents types de stockages. Plusieurs fonctionnalités peuvent être offertes selon le produit sélectionné. Parmi les plus courantes, nous retrouvons notamment : la protection temps réelle, l'analyse du système (automatique ou sur demande), la protection contre le vol d'identité, ou encore l'analyse de courriels. La section qui suit met l'accent sur les principales technologies et méthodes de détection utilisées. Nous présentons de plus un survol de l'industrie et du marché des produits antivirus.

### 2.2.1 Technologies et méthodes de détection

#### Signatures

Cette méthode de détection consiste à analyser les fichiers présents sur un système afin de détecter la présence de signatures associées à des fichiers malveillants connus. Par signature, nous entendons une partie du code, ou la totalité, qui permettrait l'identification du logiciel malveillant. Ce principe de détection repose sur l'utilisation d'une base de données de signatures, laquelle doit être mise à jour régulièrement afin de garantir la protection du système contre les nouvelles menaces.

#### Heuristiques

Le mot heuristique provient du mot Grec « heuriskein », qui signifie découvrir. L'analyse heuristique décrit une méthode qui permet de découvrir d'éventuels logiciels malveillants qui ne sont pas encore connus de l'antivirus. Son fonctionnement consiste à analyser le comportement supposé des programmes afin de déterminer si ces derniers sont malveillants ou non. Cette analyse peut soit être basée sur un système de règles de décision (*rule-based systems* en anglais), et/ou sur un système de poids (*weight-based systems* en anglais). Dans le premier cas, le système recherche des actions ou des combinaisons d'actions étant reconnues pour être malveillantes. Dans le second cas, le système associe un poids de « suspicion » à chaque règle et utilise la somme pondérée des poids afin de déterminer si un fichier est malveillant ou pas.

#### Comportements suspects

Cette technique de détection repose sur l'analyse permanente du comportement des logiciels actifs sur le système. Elle permet de détecter et de bloquer l'action d'un programme considéré comme potentiellement malveillant et de prévenir les dommages sur le système visé. Cette technique a comme principal avantage de permettre aux logiciels antivirus de détecter la présence de nouveaux logiciels malveillants.

#### Carré de sable

Similairement à l'analyse de comportements suspects, la détection par carré de sable (*sandbox* en anglais) consiste à analyser le comportement des logiciels malveillants. Alors que dans le premier cas les actions malveillantes sont détectées lors de leur exécution, la détection par carré de sable repose sur l'exécution des fichiers malveillants dans un environnement virtuel.

## Apprentissage automatique

Ces méthodes de détection sont basées sur des algorithmes d'apprentissage machine. Ces derniers peuvent être divisées en deux grandes catégories, soit l'apprentissage supervisé, et l'apprentissage non supervisé. Dans le premier cas, le modèle de détection est construit à partir d'un échantillon de fichiers étiquetés comme étant malveillants ou bénins. Le modèle apprend les caractéristiques qui distinguent les fichiers malveillants des fichiers bénins, et peut ainsi identifier si un nouveau fichier est malveillant uniquement à partir de ses caractéristiques. Dans le second cas, les fichiers ne sont pas étiquetés. Le modèle doit alors apprendre par lui-même quelle est la structure sous-jacente des données. Bien que les méthodes d'apprentissage automatique permettent de détecter de nouveaux fichiers malveillants, les modèles développés doivent tout de même être mis à jour afin de prendre en compte l'évolution continue des logiciels malveillants.

## Ressources en ligne

Les logiciels antivirus peuvent recourir à différentes ressources en ligne afin de détecter une attaque par logiciel malveillant. Le recours à des listes noires permet de bloquer, par exemple, l'exécution de programmes listés dans une liste noire. À l'inverse, une liste blanche permet uniquement l'exécution des programmes qui y sont listés.

### 2.2.2 Marché et produits antivirus

Il existe actuellement plusieurs dizaines de produits antivirus disponibles sur le marché. Parmi les principaux acteurs de l'industrie, nous retrouvons, notamment, les compagnies suivantes : Symantec, McAfee, Microsoft, ESET, Trend Micro, Avast, AVG, et Kaspersky. Alors qu'une majorité de compagnies offrent des produits payants, d'autres, comme AVG, Avast et Microsoft, offrent des solutions gratuites.

## Industrie des antivirus

La prolifération des menaces informatiques des dernières années a eu pour conséquence d'augmenter la croissance de l'industrie des produits antivirus. Au cours de l'année 2015, le marché a atteint 22.1 milliard de dollars US, représentant une importante partie du marché global de la sécurité des systèmes d'information (Statista, 2015). En fait, l'utilisation des logiciels antivirus est considérée comme la solution de protection la plus utilisée (AV Comparatives, 2013).

Selon une étude réalisée en 2012 par McAfee auprès de plusieurs millions d'utilisateurs (McAfee Labs, 2012), 83% des ordinateurs personnels sont protégés par un produit de sécurité, tel qu'un logiciel antivirus, un logiciel anti-espion ou un pare-feu. Ces chiffres baissent toutefois à 76% si nous prenons en compte uniquement les systèmes (Windows) qui ont une protection en temps réel activée et à jour (Microsoft Corporation, 2014a). En d'autres mots, environ 24% des systèmes n'ont pas de solution antivirus installée, ou ont une solution de protection qui n'est plus à jour, ou désactivée. Ces statistiques sont toutefois fonction de différents facteurs, tels que la région, ou le système d'exploitation installé sur le système. Par exemple, alors que plus de 50% des systèmes Windows 7 et Windows Vista n'ont pas de logiciel antivirus installé en 2017, moins de 5% des systèmes Windows 10 sont non protégés (Corporation, 2017). Ces différences sont aussi observable au niveau géographique ; 73% des systèmes Windows situés au Pérou sont protégés par une solution en temps réel contre 92% pour la Finlande.

## 2.3 Santé publique

Dans la section qui suit, nous présentons en un premier temps un bref aperçu de la santé publique, de son historique, ainsi que des principaux concepts de base associés à cette dernière. En un second temps, nous énonçons les principales fonctions essentielles de la santé publique, ainsi que son application en termes de cadre de travail et modèle d'implémentation.

### 2.3.1 Définitions et historique

#### Définitions

Bien que le concept de santé publique soit largement utilisé, il n'existe pas de définition simple et univoque. La majorité des définitions modernes de la santé publique remontent à celle de Winslow (1920), qui proposa en 1920 la définition suivante :

*“La santé publique est la science et l'art de prévenir les maladies, de prolonger la vie et de promouvoir la santé et l'efficacité physiques à travers les efforts coordonnés de la communauté pour l'assainissement de l'environnement, le contrôle des infections dans la population, l'éducation de l'individu aux principes de l'hygiène personnelle, l'organisation des services médicaux et infirmiers pour le diagnostic précoce et le traitement préventif des pathologies, le développement des dispositifs sociaux qui assureront à chacun un niveau de vie adéquat pour le maintien de la santé, l'objet final étant de permettre à chaque individu de jouir de son droit inné à la santé et à la longévité. ”*

Plus d'un siècle suivant, cette description –ainsi que les nombreuses autres qui en ont été

dérivées– ont donné lieu à de nombreux débats sur la définition et la pratique de la santé publique. Somme toute, le concept de santé publique est principalement utilisé avec deux sens : l’un servant à distinguer les services publics et privés, et l’autre permettant de séparer l’aspect individuel de l’aspect collectif.

## **Courants historiques**

Les origines de la santé publique sont généralement situées au XVIII<sup>e</sup> siècle et à la période préindustrielle. En particulier, quatre grands courants ont façonné la santé publique telle que nous la connaissons aujourd’hui : 1) l’hygiène publique, 2) l’hygiène personnelle, 3) la médecine sociale, et 4) l’organisation des services de santé. L’hygiène publique est probablement le courant le plus ancien, et celui auquel la santé publique est la plus fréquemment identifiée. Remontant au Moyen-Age, l’apparition de plusieurs maladies infectieuses, telles que la peste et la lèpre, a donné lieu à plusieurs mesures publiques de contrôle des épidémies. Les travaux de Pasteur en bactériologie ont par la suite ouvert la voie à une “nouvelle santé publique” à la fin du XIX<sup>e</sup> siècle. C’est notamment à partir du XIX<sup>e</sup> siècle que de grandes campagnes sanitaires sont organisées afin de promouvoir de bonnes pratiques d’hygiène personnelle. Le milieu du XX<sup>e</sup> siècle est alors marqué par une diminution des maladies infectieuses. Cependant, l’augmentation des maladies chroniques et dégénératives amène au développement d’études épidémiologiques sur les comportements à risque, tels que l’alcoolisme, le tabagisme, ou le manque d’activité physique. Le milieu du XIX<sup>e</sup> siècle est par la suite marqué par une reconnaissance que les conditions sociales et économiques ont des effets importants sur la santé. Cette reconnaissance a entre autres conduit à une responsabilisation de l’État quant à la santé de la population et de certains groupes vulnérables. Plusieurs interventions directes de l’État sont alors implémentées, donnant le jour au concept de médecine sociale. La fin du XIX<sup>e</sup> siècle marque le développement des systèmes de soins de santé, et le rôle prépondérant de l’État. Notamment, la première intervention de l’État, qui préfigure les systèmes d’assurance-maladie d’aujourd’hui, fait son apparition en Allemagne. S’en suit alors un développement important des soins de santé à partir de la première moitié du XX<sup>e</sup> siècle.

### **2.3.2 Déterminants de la santé**

Les déterminants de la santé peuvent être définis comme l’ensemble des facteurs qui influencent l’état de santé d’une population. Ces facteurs peuvent être reliés, entre autres, aux caractéristiques individuelles (comportement, génétique, âge), aux milieux de vie (familial, scolaire, communautaire), aux systèmes (éducation, santé, solidarité sociale), ou encore au contexte global (économie, culture, politique). À titre d’illustration, les déterminants suivants

sont reconnus par l'Agence de la santé publique du Canada (Agence de la santé publique du Canada, 2011) :

1. le niveau de revenu et le statut social ;
2. les réseaux de soutien social ;
3. l'éducation ;
4. l'emploi et les conditions de travail ;
5. les environnements sociaux ;
6. les environnements physiques ;
7. les habitudes de santé et la capacité d'adaptation personnelle ;
8. le développement de la petite enfance ;
9. le patrimoine biologique et génétique ;
10. les services de santé ;
11. le sexe ;
12. la culture.

Un déterminant peut soit être une *cause directe* de l'état de santé, ou encore une *cause indirecte*. Dans le premier cas, le déterminant agit directement sur l'état de santé d'une population alors que dans le second, le déterminant agit sur la santé par un ensemble de facteurs dits intermédiaires. Selon le niveau de proximité du déterminant, ce dernier peut être soit qualifié de *facteur distal*, *facteur intermédiaire*, ou *facteur proximal*. Alors qu'un facteur proximal agit directement ou presque directement sur la santé, un facteur distal se situe plus loin dans la chaîne causale et agit par un ensemble de facteurs intermédiaires. Bien que certains déterminants puissent agir de manière isolée sur la santé, leur effet est souvent le résultat d'interactions complexes. De plus, tout déterminant se situe dans le temps et l'espace. L'effet d'un déterminant peut ainsi être amené à évoluer et changer en direction et magnitude selon le lieu et le temps.

### **Cadres relatifs aux déterminants de la santé**

Les multiples relations entre déterminants et l'état de santé d'une population peuvent être illustrées par l'utilisation de cadres, ou représentations visuelles. Selon l'objectif et l'audience visée, un cadre peut être catégorisé selon trois types : *cadre explicatif*, *cadre interactif* et *cadre axé sur l'action* (Conseil canadien des déterminants de la santé, 2015).



**Cadre explicatif** Un cadre explicatif vise à représenter et expliquer les différents déterminants de la santé. Bien que ce type de cadre permette de préciser la contribution relative des déterminants, il ne vise pas à illustrer les relations causales entre ces derniers et la santé.

**Cadre interactif** Aussi appelé *cadre conceptuel*, ce type de cadre permet d'illustrer les relations et les multiples liens entre les déterminants de la santé. Contrairement aux cadres explicatif, les cadres interactifs aident à visualiser et identifier les relations causales entre les déterminants et la santé. Bien que ces cadres peuvent aider à guider l'élaboration de politiques ou encore élaborer un plan de recherche, ils ne permettent pas de définir des stratégies d'interventions.

**Cadre axé sur l'action** Ce type de cadre, aussi nommé *cadre d'action*, permet de soutenir la prise de décisions. Il peut aider à guider les décideurs politiques, les chercheurs et les praticiens à cibler et à développer des interventions visant à améliorer l'état de santé.

En fonction de ses caractéristiques, un cadre peut appartenir à une ou plusieurs de ces catégories. Un cadre peut aussi être défini par sa portée, qui peut être soit *élargie* ou *étroite*. Alors que le premier type est axé sur les déterminants d'une population entière, le second est limité à une sous-population. À titre d'exemple, le modèle développé par Dahlgren et Whitehead (1991) est le plus connu et le plus utilisé (voir Figure 3.1). Ce modèle élargi des déterminants sociaux de la santé est basé sur une approche holistique, c'est-à-dire qu'il prend en compte plusieurs dimensions liées à la santé, telles que l'économie, l'éducation, l'environnement ou l'individu.

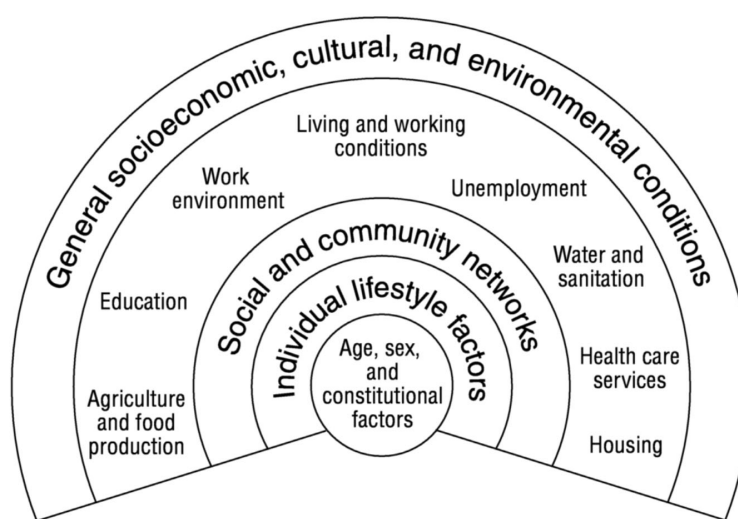


Figure 2.1 Modèle élargi des déterminants sociaux de la santé (Dahlgren et Whitehead, 1991)

### 2.3.3 Fonctions essentielles

L'approche axée sur la santé publique repose principalement sur quatre fonctions essentielles : la promotion de la santé, la protection de la santé, la prévention, et la surveillance (Ministry of Health Services Province of British Columbia, 2005).

**Promotion** L'objectif principal de cette fonction est de promouvoir la santé et le bien-être au sein de la population. Cette fonction désigne toute mesure visant à encourager les comportements sains et à conférer aux populations un plus grand pouvoir sur leur santé.

**Protection** La fonction de protection regroupe l'ensemble des interventions menées afin de limiter et réduire les risques pour la santé d'une population. Ces mesures visent principalement les grands déterminants de la santé plutôt que les facteurs au niveau individuel.

**Prévention** La prévention désigne l'ensemble des interventions visant à éviter ou réduire l'incidence et la gravité des maladies ou des accidents. Les différentes stratégies de prévention en santé peuvent être classifiées en trois catégories : la *prévention primaire*, la *prévention secondaire*, et la *prévention tertiaire*. La prévention primaire désigne les interventions qui visent à réduire ou limiter les risques d'apparition de maladies. Ce type de prévention va cibler soit les causes spécifiques ou les facteurs de risque associés à certaines maladies afin d'en prévenir le développement. Les mesures de prévention secondaire comprennent les méthodes de détection des maladies afin d'en limiter la propagation. Une fois que la maladie est développée, la prévention tertiaire veille à réduire l'impact de la maladie et améliorer la qualité de vie.

**Surveillance** La fonction de surveillance vise à observer et mesurer de manière continue l'état de santé de la population ainsi que de ses déterminants. L'exercice de cette fonction permet, entre autres, de dresser un portrait global de la santé de la population, d'observer des tendances et des variations temporelles et spatiales, et de suivre l'évolution de certains problèmes spécifiques de santé.

### 2.3.4 Cadre de travail

Tel qu'illustré à la Figure 2.2, le cadre de travail utilisé en santé publique peut être divisé en quatre étapes principales, soit : 1) définition du problème ; 2) identification des facteurs de protection et des facteurs de risque ; 3) développement et évaluation de stratégies et d'interventions et 4) promotion afin d'assurer une adoption massive par la population.

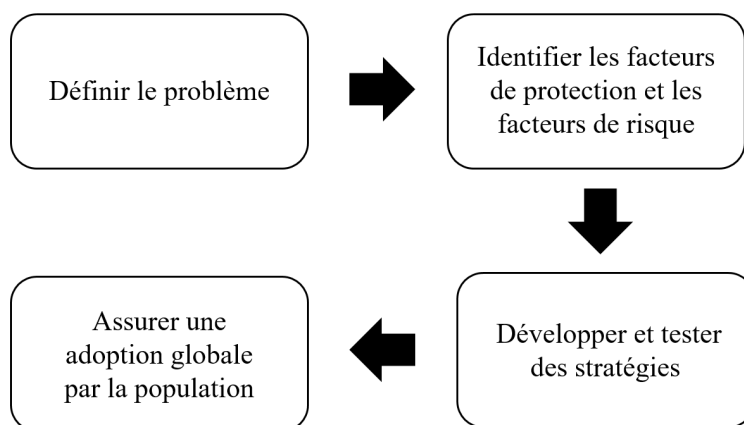


Figure 2.2 Modèle de prévention en santé publique

**Définition du problème** La première étape, soit la définition du problème, consiste à identifier et comprendre le “qui”, “quoi”, “quand”, “où”, et “comment”. Similaire à la fonction de surveillance, cette étape implique d’analyser des données afin de mieux comprendre la fréquence du problème, son emplacement et ses tendances.

**Identification des facteurs** Une fois que la population est connue et que le problème est ciblé, il convient d’identifier quels sont les causes et les corrélats du problème. Plus spécifiquement, cette étape consiste à identifier les facteurs de risque et les facteurs de protection qui sont associés au problème. Un facteur de risque contribue à augmenter la probabilité d’occurrence du problème. À l’opposé, un facteur de protection contribue à diminuer la probabilité. Connaître ces facteurs est particulièrement important afin de mieux orienter les stratégies de prévention.

**Développement et évaluation de stratégies** Cette étape consiste à développer des stratégies de prévention à partir des connaissances existantes et des résultats obtenus lors des deux étapes précédentes. Les cibles d’interventions visées par les stratégies développées peuvent être les individus, l’environnement interpersonnel, les organisations, les communautés, ou encore les politiques publiques. Une fois qu’une stratégie est développée, la dernière étape consiste à évaluer cette dernière afin de déterminer son efficacité réelle.

**Implémentation et adoption** Lors de cette étape, les stratégies de prévention qui se sont révélées efficaces sont implémentées et disséminées auprès de la population afin d’en assurer une adoption massive. L’impact et le rapport coût-efficacité de chaque stratégie sont par la suite continuellement évalués.

### 2.3.5 Modèles d'implémentation

Nous présentons dans la section suivante les principaux modèles d'implémentation utilisés en santé publique, soit : 1) le modèle à croyances pertinentes à la santé, 2) le modèle des étapes du changement, 3) le modèle socio-écologique, et 4) le modèle PRECEDE-PROCEED.

**Modèle de croyances pertinentes à la santé** Ce modèle psychologique est l'un des plus anciens et des mieux connus en terme de modélisation du comportement relié à la santé. Il repose sur le fait que trois principes amènent une personne à adopter, ou non, une recommandation visant à modifier son comportement : 1) la perception de la maladie (risque et gravité), 2) la perception des avantages et des contraintes reliés au changement de comportement, et 3) les incitatifs pour passer à l'action

**Modèle des étapes du changement** Développé à la base pour la cessation du tabagisme, ce modèle vise à représenter en six étapes le processus par lequel les individus passent pour effectuer un changement durable d'attitude et de comportement. Ces étapes sont : 1) la préreflexion, 2) la réflexion, 3) la décision, 4) l'action, 5) le maintien, et 6) l'intégration. La compréhension de ces six étapes est particulièrement intéressante au moment d'élaborer une stratégie de communication visant à changer le comportement d'une population.

**Modèle socio-écologique** Le modèle socio-écologique est considéré comme étant une composante clef de la santé publique moderne. Ce modèle permet de définir un cadre de recherche et d'intervention axé sur une vision élargie des déterminants de la santé. En d'autres mots, ce modèle intègre à la fois les facteurs interpersonnels, intrapersonnels, et socio-environnementaux. Son application est particulièrement utile dans un contexte d'identification de déterminants, de développement et d'évaluation de stratégies, et de promotion.

**Modèle PRECEDE-PROCEED** Ce modèle de planification est basé sur les disciplines de l'épidémiologie, des sciences sociales, et de l'éducation. L'acronyme PRECEDE signifie en anglais "Predisposing, Reinforcing and Enabling Constructs in Educational/Environment Diagnosis and Evaluation". L'acronyme PROCEED signifie pour sa part en anglais "Policy, Regulatory and Organizational Constructs in Educational and Environmental Development". En d'autres mots, la première composante du modèle (PRECEDE) représente l'appréciation et l'analyse des besoins, et la seconde (PROCEED) procède à la mise en application et à l'évaluation du programme qui doit répondre aux besoins identifiés. Le modèle PRECEDE-PROCEED est particulièrement utilisé pour développer et évaluer des stratégies de promotion

de la santé.

## 2.4 Sécurité et santé publique : Travaux antérieurs

L'analogie entre la santé et la sécurité des systèmes d'information n'est pas nouvelle dans la littérature de la sécurité informatique. De nombreux travaux ont d'ailleurs déjà exploré l'idée d'appliquer à la sécurité des systèmes d'information des concepts et méthodes inspirés de la santé publique. À cet effet, nous présentons brièvement dans la section qui suit les travaux antérieurs qui ont étudié le développement et l'application de concepts empruntés à la santé publique en sécurité des systèmes d'information.

### 2.4.1 Travaux de recherche

#### Modèle de gouvernance

Plusieurs travaux de recherche se sont inspirés de la santé publique pour développer des modèles ou encore des stratégies globales pouvant être appliqués au domaine de la sécurité informatique. Par exemple, Rice *et al.* (2010) ont exploré l'analogie avec la santé publique pour présenter une stratégie globale visant à protéger la “santé” du *cyberespace*. La stratégie mise en avant repose sur cinq composantes : i) désinfecter l'environnement, ii) contrôler les infections communautaires, iii) éduquer les acteurs, iv) organiser des services de détection et de prévention, et v) créer la machinerie sociale pour la santé du cyberspace. Dans un rapport publié par Microsoft, Charney (2010) propose de se baser sur les principes de la santé publique pour développer des activités visant à améliorer et maintenir la santé des appareils dans l'écosystème informatique. À titre d'exemple, l'auteur mentionne la promotion de mesures préventives, la détection des appareils infectés, la notification des utilisateurs affectés, le traitement de ces appareils, et l'adoption de mesures additionnelles afin de limiter la propagation des infections à d'autres appareils.

D'autres travaux poussent l'analogie en considérant le concept de *cybersanté* comme un bien public, au même titre que la santé. Mulligan et Schneider (2011) appellent à considérer la cybersécurité comme un bien commun et à adopter des mécanismes inspirés de ceux utilisés en santé publique. À cet effet, ils proposent une nouvelle doctrine pour la sécurité informatique : la cybersécurité publique. Cette dernière est définie comme toute doctrine en cybersécurité qui vise à produire la cybersécurité, et à gérer l'insécurité restante en considérant la balance entre les droits individuels et le bien-être public. Les auteurs proposent, notamment, de gérer l'insécurité par la surveillance, la mise à jour des logiciels ou encore l'isolation des systèmes. Similairement, Sullivan *et al.* (2012) considèrent la santé d'Internet comme un bien public

et proposent de développer un modèle basé sur la santé publique afin de protéger l'Internet. Leur modèle est, entre autres, basé sur le principe que la santé d'Internet est une responsabilité partagée, qui demande le développement d'approches basées sur des données probantes. Finalement, les auteurs discutent l'implémentation d'un tel modèle pour l'Internet et identifient cinq domaines nécessitant davantage de recherche et de développement : i) l'expérience de l'utilisateur, ii) le développement d'interventions éducatives systématiques et ciblées, iii) la division des rôles et des responsabilités au sein des différentes entités, iv) l'établissement de métriques, mesures et de systèmes de partage d'information, et v) l'évaluation des politiques et des technologies pour combattre les logiciels malveillants et promouvoir une bonne hygiène informatique.

## Prévention des attaques

D'autres chercheurs se sont concentrés sur l'application de concepts empruntés à la santé publique dans un contexte spécifique de prévention. Par exemple, Rowe *et al.* (2012a) ont développé un cadre de travail pour identifier et définir certaines menaces informatiques ainsi que des solutions potentielles. Plus spécifiquement, ils proposent d'appliquer leur cadre de travail pour mieux comprendre les préférences de risque individuel en cybersécurité, le tout dans le but d'identifier les stratégies d'interventions qui seront plus efficaces à prévenir les cyberattaques.

À partir de ces travaux, Rowe *et al.* (2012b) ont publié un rapport détaillé où ils explorent les similarités entre les vaccins et les solutions antivirus. En s'inspirant de la recherche sur la préférence de risques reliés aux vaccins, les auteurs développent un sondage visant à analyser les préférences reliées aux solutions antivirus et à la perception individuelle des menaces en cybersécurité. Les résultats du sondage ont indiqué un certains nombres de différences intéressantes entre les préférences reliées aux vaccins et aux solutions antivirus, confirmant les bénéfices potentielles d'une approche basée sur la santé publique pour la cybersécurité.

Dans sa thèse intitulée "Avoiding the cyber pandemic : A Public Health Approach to Preventing Malware Propagation", Zelonis (2004) propose de s'inspirer de la santé publique pour prévenir la propagation des logiciels malveillants. En explorant l'analogie entre le Syndrome d'immunodéficience acquise (SIDA) et les logiciels malveillants, l'auteur propose des stratégies qui combinent une approche comportementale et technologique afin de modifier les comportements à risque qui contribuent à faciliter la propagation des logiciels malveillants.

## Surveillance et partage de données

Certains travaux de recherche se sont concentrés sur les aspects reliés à la fonction de surveillance en sécurité des systèmes d'information. Notamment, Sedenberg et Mulligan (2015) élaborent sur comment les principes de santé publique peuvent guider le partage d'information en sécurité des systèmes d'information. À cet effet, les auteurs dérivent quatre principes de base et discutent comment ces derniers peuvent s'appliquer dans différents mécanismes de partage de données. Similairement, Parker et Farkas (2011) proposent la création d'un système de surveillance en sécurité des systèmes d'information. À partir des données recueillies par ce système, les chercheurs proposent de développer des modèles statistiques qui permettraient d'identifier et de quantifier les facteurs de risque reliés aux attaques informatiques, et d'estimer le risque de succès des attaques.

### 2.4.2 Implémentation et initiatives existantes

Alors que les travaux précédents ont principalement discuté des implications d'un modèle basé sur la santé publique pour la sécurité des systèmes d'information, d'autres se sont concentrés sur les aspects techniques reliés à l'implémentation d'un tel modèle. Notamment, Rowe *et al.* (2013) décrivent certaines des caractéristiques techniques et opérationnelles reliées à l'implémentation d'un modèle de santé publique pour la cybersécurité. Plus spécifiquement, ils élaborent les fonctions de surveillance, prévention, et de réponse en cas d'incident. Par exemple, ils proposent de définir et de surveiller en ligne des indicateurs de cybersanté, tels que des mots clés ou des *hashtag*, afin de détecter la présence d'une nouvelle attaque.

Fryer (2012) explore l'analogie entre la santé publique et la sécurité des systèmes d'information. En particulier, il s'intéresse au cas des logiciels malveillants qui se propagent par les sites Web, soit par le téléchargement furtif. Il développe un cadre de travail basé sur l'approche de la santé publique pour prévenir les attaques Web et propose des interventions. Plus spécifiquement, il fait une simulation pour montrer l'efficacité d'une intervention.

Plus récemment, l'équipe de réponse d'urgence informatique du Japon, la *Japan Computer Emergency Response Team* (JPCERT) en anglais, a introduit la Cyber Green Initiative (CGI) (JPCERT Coordination Center, 2014, 2015). Dans un rapport publié en 2015, il est proposé d'appliquer les leçons apprises en santé publique afin d'améliorer la santé globale du *cyberécosystème*. Plus spécifiquement, le rapport appelle à une approche de collaboration globale entre les différents intervenants afin i) d'établir une plate-forme pour générer des statistiques fiables et comparables sur la cybersanté, ii) de permettre des efforts opérationnels d'assainissement, et iii) de fournir des informations sur les risques systémiques du cyberécosystème.

## CHAPITRE 3 MODÈLE DE PRÉVENTION

Ce chapitre présente le modèle de prévention des attaques par logiciels malveillants appliqué dans le cadre de la présente thèse. Nous y introduisons en un premier temps un cadre de travail applicable à la sécurité des systèmes d'information, plus particulièrement aux attaques par logiciels malveillants. En un second temps, nous élaborons sur les détails de son application dans le contexte spécifique de la prévention des attaques par logiciels malveillants. Finalement, nous discutons des travaux proposés en terme d'originalité et de contributions.

### 3.1 Cadre de travail

La section qui suit porte sur le cadre de travail développé pour mener à bien notre projet de recherche. Plus spécifiquement, nous présentons les étapes nécessaires à l'atteinte de la première partie de notre objectif général, soit le *développement d'un modèle de prévention basé sur l'approche de la santé publique* pour la prévention des attaques par logiciels malveillants.

À cet effet, le modèle de prévention développé est fortement inspiré de l'approche utilisée en santé publique (voir Figure 2.2). Similairement, notre modèle peut être divisé en quatre étapes, soit : 1) définition du problème ; 2) identification des déterminants, soit des facteurs de protection et des facteurs de risque ; 3) développement et évaluation de stratégies, et 4) implémentation et promotion des stratégies.

Tableau 3.1 Modèle de prévention des attaques par logiciels malveillants

Étape	Description	Application
Définition du problème	Identifier et comprendre le “qui”, “quoi”, “quand”, “où”, et “comment”.	Étudier la prévalence et les patrons d'agrégation des attaques par logiciels malveillants.
Identification des déterminants	Identifier quels sont les causes et les corrélats du problème.	Identifier les déterminants qui sont associés aux attaques par logiciels malveillants.
Développement et évaluation de stratégies	Développer et évaluer des stratégies de prévention.	Développer et évaluer des stratégies visant à prévenir et/ou réduire l'occurrence des attaques par logiciels malveillants.
Implémentation et promotion des stratégies	Implémenter et disséminer auprès de la population les stratégies.	Implémenter et promouvoir les stratégies de prévention des attaques par logiciels malveillants qui ont été prouvées efficaces.



### 3.1.1 Définition du problème

Tel que présenté dans le Tableau 3.1, cette étape consiste à étudier la prévalence et les patrons d’agrégation des attaques par logiciels malveillants afin de mieux comprendre leur fréquence, leurs emplacements et leurs tendances. Cette étape est similaire à la discipline de renseignement sur les menaces, *threat intelligence* en anglais. Autrement dit, elle revient à collecter et organiser des informations liées aux attaques informatiques afin de surveiller le portrait global du phénomène et des tendances. Il existe notamment plusieurs entreprises du secteur privé qui se spécialisent en *threat intelligence*. Par exemple, il est fréquent pour les compagnies antivirus de publier des statistiques sur la prévalence et les tendances des attaques informatiques observées chez leurs clients. Cependant, définir la pleine magnitude du problème des attaques par logiciels malveillants est principalement entravé par le manque d’informations systématiques, et de données probantes accessibles et appropriées. Un exemple de solution consiste à déployer des ressources afin de collecter, analyser, et interpréter des données d’attaques afin de produire des rapports visant à améliorer la compréhension du problème. À titre d’exemple, il convient de mentionner à nouveau l’initiative de l’équipe de réponse d’urgence informatique du Japon, la CGI, qui vise entre autres à générer des statistiques fiables et comparables sur la cybersanté et fournir des informations sur les risques systémiques du cyberécosystème (JPCERT Coordination Center, 2014, 2015).

### 3.1.2 Identification des déterminants

Cette seconde étape revient à identifier quels sont les déterminants reliés aux attaques par logiciels malveillants. En d’autres mots, cela revient à étudier les facteurs de risque et les facteurs de protection qui sont les causes et les corrélats du problème. Dans une optique de santé publique, les facteurs multi-niveaux (systèmes, usagers, environnement, etc.) sont examinés afin de mieux comprendre leur impact et leurs interactions. Par exemple, l’analyse des données collectées par certaines entreprises privées en sécurité a permis d’identifier certains systèmes d’exploitation, ou régions géographiques, qui sont exposés à un risque plus élevé d’attaques par logiciels malveillants (voir Section 2.1.4). On reconnaît en général que les variations entre systèmes d’exploitation sont attribuables en grande partie à la popularité et au nombre de vulnérabilités de certains systèmes. Or, nous comprenons beaucoup moins bien l’impact du contexte géographique, socio-environnemental et politique sur le risque d’attaques. Les lacunes en matière de données de surveillance des attaques, soit l’étape de définition du problème, viennent par conséquent limiter la compréhension des corrélats, et particulièrement des causes, qui influent sur le risque d’attaque par logiciels malveillants.

### 3.1.3 Développement et évaluation de stratégies

La troisième étape du modèle de prévention concerne le développement et l'évaluation de stratégies. Ces dernières se doivent d'être multidisciplinaires, et de travailler à l'échelle des systèmes, des individus, des populations, et de l'environnement. À l'instar de la santé, les stratégies peuvent être classifiées en trois niveaux : primaire, secondaire, et tertiaire. La prévention primaire vise à protéger les systèmes et ses utilisateurs, et à prévenir les infections par logiciels malveillants. À titre d'exemples, nous retrouvons les stratégies visant à prévenir l'adoption de comportement usager à risque en ligne, favoriser la mise à jour régulière des systèmes, ou encore le recours à un pare-feu. La prévention secondaire, quant à elle, cherche à détecter les infections le plus tôt possible afin d'en ralentir ou d'en arrêter la progression. Un exemple notable à ce niveau serait le recours à un logiciel antivirus. Et la prévention tertiaire désigne les stratégies reliées à l'arrêt de la progression des infections, et le contrôle des répercussions non désirées. Ceci inclut notamment la discipline de réponse aux incidents. Bien qu'il existe plusieurs efforts reliés au développement de nouvelles stratégies de prévention en sécurité des systèmes d'information, peu d'efforts sont toutefois dédiés à leur évaluation et leur adoption. Avec une approche axée sur la santé publique, l'évaluation des stratégies implique non seulement d'évaluer l'efficacité en terme de réduction des attaques par logiciels malveillants, mais en plus les retombées économiques et les incidences à long terme. Ces nouvelles données collectées peuvent ainsi servir à la prise de décisions afin de soutenir les efforts de prévention.

### 3.1.4 Implémentation et promotion des stratégies

La dernière étape du modèle de prévention consiste à implémenter et promouvoir les stratégies de prévention des attaques par logiciels malveillants qui ont été prouvées efficaces. Tel que mentionné à l'étape précédente, l'évaluation des stratégies doit se poursuivre au cours de l'implémentation de telle sorte à évaluer l'impact économique, et les effets à long terme. Quant à la promotion des stratégies, elle repose principalement sur une communication efficace auprès de la population. Notamment, les informations transmises doivent être facilement comprises et publiées par un vaste éventail de médias et de canaux de communication. Suivant les principes de la santé publique, les objectifs visés sont d'engager la population, et d'établir une mesure de confiance avec cette dernière à l'égard de la stratégie proposée. Pour ce faire, il convient d'impliquer plusieurs organisations (secteur privé et public) afin d'assurer une coordination adéquate quant au partage de l'information. À cet effet, le réseau canadien SERENE-RISC représente un exemple d'initiative canadienne qui regroupe le secteur privé et public afin de communiquer, entre autres, les meilleures pratiques reliées à la sécurité des

systèmes d'information (SERENE-RISC, 2018). Le mois de la sensibilisation à la cybersécurité est un autre exemple notable. Cette campagne de portée internationale vise notamment à informer le grand public de l'importance de la sécurité des systèmes d'information, et de mesures à suivre pour se protéger en ligne (National Cyber Security Alliance, 2018).

### 3.2 Application aux attaques par logiciels malveillants

La présente section porte sur l'application du modèle de prévention précédemment développé. En particulier, cette section se concentre sur la seconde partie de notre objectif général, soit *l'application d'un modèle basé sur l'approche de la santé publique pour la prévention des attaques par logiciels malveillants*.

L'application du modèle développé sera axée sur la deuxième et troisième étape du Tableau 3.1. La réalisation de la deuxième étape s'inscrit dans le cadre de notre premier objectif spécifique, soit *l'identification des causes et des corrélats reliés aux attaques par logiciels malveillants*. Nous tenterons ainsi d'identifier des facteurs de risque et des facteurs de protection qui sont associés aux attaques par logiciels malveillants. Cette étape sera basée sur une approche écologique, c'est-à-dire que nous prendrons en compte plusieurs dimensions, telles que les facteurs socio-économiques, politiques et individuels. Les détails de l'implémentation sont présentés dans la Section 3.2.1. La troisième étape est pour sa part reliée à notre second objectif spécifique, soit *l'évaluation de l'efficacité réelle d'une intervention visant à prévenir et/ou réduire l'occurrence des attaques par logiciels malveillants*. Plus spécifiquement, l'étape trois sera axée sur l'évaluation d'une méthode de prévention déjà existante; les solutions antivirus. La Section 3.2.2 présente les détails de l'implémentation.

#### 3.2.1 Déterminants des attaques par logiciels malveillants

*Quels sont les causes et les corrélats reliés aux attaques par logiciels malveillants ?*

La section qui suit vise en un premier temps à présenter une synthèse des travaux de recherche sur l'identification des déterminants reliés aux attaques par logiciels malveillants. En un second temps, nous développons un modèle écologique des déterminants reliés aux attaques par logiciels malveillants à partir de la littérature existante. Finalement, nous exposons et mettons en contexte les travaux de recherche proposés dans le cadre de la présente thèse.

## Synthèse des travaux existants

Notre synthèse est basée sur les travaux portant sur les logiciels malveillants en combinaison avec l'un des termes suivants : facteur(s) de risque (*risk factor(s)* en anglais), facteur(s) de protection (*protective factor(s)* en anglais), ou déterminant(s) (*determinant(s)* en anglais). Plus spécifiquement, nous nous sommes intéressés aux travaux qui ont étudié la relation entre les attaques par logiciels malveillants et différents facteurs de risque et de protection. Les travaux rapportant uniquement des statistiques descriptives sur la prévalence ou la fréquence des attaques par logiciels malveillants, ainsi que les travaux de nature théorique ne sont pas couverts. Au total, 26 articles de recherche ont été recensés. Ces derniers ont été classifiés en deux catégories principales selon la population étudiée. La première contient 11 articles qui se sont concentrés sur l'étude des attaques par logiciels malveillants au niveau des pays et des organisations. La seconde catégorie contient 15 articles qui ont étudié les facteurs au niveau des systèmes et des individus. Pour chaque catégorie, nous présentons l'indicateur d'attaques par logiciels malveillants qui a été utilisé, la population ciblée par l'étude, et la nature des déterminants qui ont été analysés. Les Tableau 3.2 et Tableau 3.3 présentent respectivement les articles de la première catégorie et de la seconde catégorie.

**Niveau institutionnel et organisationnel** Alors que la grande majorité des articles ont choisi d'étudier les attaques par logiciels malveillants au niveau des pays, seulement trois articles sur 11 se sont concentrés sur l'analyse des facteurs de risque et de protection au niveau des organisations. Quant aux indicateurs d'attaques, cinq études ont utilisé les attaques bloquées (détections), et six les attaques réussies (infections). Finalement, parmi l'ensemble des travaux, deux se sont concentrés sur un type d'attaque en particulier ; Van Eeten *et al.* (2010) ont étudié la prévalence du spam au niveau des pays et des fournisseur d'accès Internet (FAI), et Asghari *et al.* (2015) ont analysé la prévalence du logiciel malveillant Conficker au niveau des pays.

**Niveau individuel** Près de la moitié des travaux réalisés au niveau individuel sont basés sur des indicateurs d'attaques auto-rapportées par les usagers. À l'opposé, l'autre moitié est basée sur des données réelles d'attaques bloquées (détections) ou réussies (infections). De plus, la grande majorité des travaux, soit 13 sur 15, ont étudié les attaques par logiciels malveillants uniquement au niveau des systèmes ou des usagers. Autrement dit, sur l'ensemble des 15 articles, uniquement deux études ont considéré l'utilisateur et le système comme un tout dans leur analyse. Parmi les 10 études au niveau usager, sept sont limitées à une population particulière définie soit par un facteur géographique ou une caractéristique individuelle.

Tableau 3.2 Indicateurs d'attaques, population et déterminants étudiés (niveau institutionnel et organisationnel)

Étude	Indicateur d'attaques	Population	Déterminants
Png <i>et al.</i> (2008)	Attaques bloquées	Pays	politiques nationales
Van Eeten <i>et al.</i> (2010)	Attaques réussies (spam)	Pays, FAI	niveau d'éducation, développement économique, piratage logiciel, développement des TIC, politiques nationales, FAI (taille, prix, technologie, part de marché)
Kleiner <i>et al.</i> (2013)	Attaques réussies	Pays	politiques nationales
Garg <i>et al.</i> (2013)	Attaques réussies	Pays	gouvernance, cadre légal, ressources économiques, sécurité des infrastructures des TIC, disponibilité des TIC, expertise nationale en sécurité de l'information
Burt <i>et al.</i> (2014)	Attaques réussies	Pays	accès numérique, stabilité institutionnelle, développement économique
Mezzour <i>et al.</i> (2015)	Attaques bloquées	Pays	développement économique, développement des TIC, expertise nationale en sécurité de l'information, relations internationales, piratage logiciel, navigation web
Asghari <i>et al.</i> (2015)	Attaques réussies (Conflicker)	Pays	initiatives anti-botnet, développement et disponibilité des TIC, piratage logiciel, part de marché des systèmes d'exploitation
Subrahmanian <i>et al.</i> (2016)	Attaques bloquées	Pays	développement économique, développement et disponibilité des TIC, piratage logiciel, fichiers binaires téléchargés/installés

Tableau 3.2 Indicateurs d'attaques, population et déterminants étudiés (niveau institutionnel et organisationnel)

Étude	Indicateur d'attaques	Population	Déterminants
Yen <i>et al.</i> (2014)	Attaques bloquées	Entreprises	pays, activité réseaux, navigation web, démographie et caractéristiques usager (genre, type d'emploi)
Thonnard <i>et al.</i> (2015)	Attaques bloquées	Entreprises	entreprise (taille, secteur d'activité), employés (type et niveau d'emploi, localisation, connexions LinkedIn)
Kumar <i>et al.</i> (2017)	Attaques réussies	Entreprises (Inde)	employé (connaissances et sensibilisation aux menaces informatiques, négligence, activité réseaux), logiciels de protection

Tableau 3.3 Indicateurs d'attaques, population et déterminants étudiés (niveau individuel)

Étude	Indicateur d'attaques	Population	Déterminants
Carlinet <i>et al.</i> (2008)	Infections	Individus/Systèmes	système d'exploitation, activité réseaux
Choi (2008)	Infections (auto-rapportées)	Individus (étudiants)	logiciels de protection (anti-virus, pare-feu, anti-logiciels espions), activité réseaux, navigation web
Bossler et Holt (2009)	Infections (auto-rapportées)	Individus (étudiants)	âge, genre, ethnicité, statut d'emploi
Ngo et Paternoster (2011)	Infections (auto-rapportées)	Individus (étudiants)	âge, genre, ethnicité, état civil
Maier <i>et al.</i> (2011)	Infections	Individus (Europe, Inde, États-Unis)	hygiène informatique (anti-virus, mise à jour logiciel), navigation web

Tableau 3.3 Indicateurs d'attaques, population et déterminants étudiés (niveau individuel)

Étude	Indicateur d'attaques	Population	Déterminants
Lee (2012)	Attaques bloquées	Individus (chercheurs académiques)	domaine d'expertise
Wilsem (2013)	Infections (auto-rapportées)	Individus (Hollande)	maîtrise de soi, activité réseaux, navigation web, logiciels de protection (anti-virus, pare-feu, anti-logiciels espions, filtre anti-spam), démographie et caractéristique usager (âge, genre, revenu, niveau d'éducation, taille et revenu du ménage, état civil, urbanisme)
Canali <i>et al.</i> (2014)	Attaques bloquées	Individus	navigation web
Vasek et Moore (2014), Vasek <i>et al.</i> (2016)	Infections	Serveurs web	type de serveur, système de gestion de contenu, mise à jour logiciel, pays d'hébergement
Leukfeldt (2015), Leukfeldt et Yar (2016)	Infections (auto-rapportées)	Individus	revenu, navigation web, système d'exploitation, navigateur web, mise à jour logiciel, anti-virus, sensibilisation aux menaces informatiques, connaissances informatiques
Ovelgönne <i>et al.</i> (2017)	Attaques bloquées	Individus/Systèmes	navigation web, fichiers binaires téléchargés/installés, type de profil utilisateur
Kranenbarg <i>et al.</i> (2017)	Infections (auto-rapportées)	Individus (Hollande)	maîtrise de soi, activité réseaux, navigation web, compétences informatiques, âge, genre, situation financière, abus de substances, état civil
Bilge <i>et al.</i> (2017)	Attaques bloquées	Systèmes	fichiers binaires téléchargés/installés, mise à jour logiciel, vulnérabilités

## Modèle écologique des déterminants

Les différents facteurs étudiés dans la section précédente ont été classifiés en trois grande catégories : i) système, ii) utilisateur, et iii) environnement et politiques (voir Figure 3.1). Les facteurs au niveau du système peuvent être d'ordre logiciel ou encore matériel. Un exemple de facteur logiciel serait le système d'exploitation d'un ordinateur, alors que l'architecture de l'ordinateur serait un facteur de type matériel. Quant au niveau utilisateur, ce dernier inclut les facteurs socio-démographiques, les caractéristiques, et les facteurs comportementaux. Pour sa part, le niveau environnement et politiques couvre autant l'environnement social, l'économie, la technologie, la gouvernance, etc. Bien que ce modèle élargi présente une division des facteurs par niveau, il est important de noter que ces derniers sont interconnectés et qu'ils n'agissent pas de manière isolée. Par exemple, une caractéristique de l'utilisateur pourrait influencer son comportement en ligne, qui lui-même viendrait modifier l'état logiciel du système.

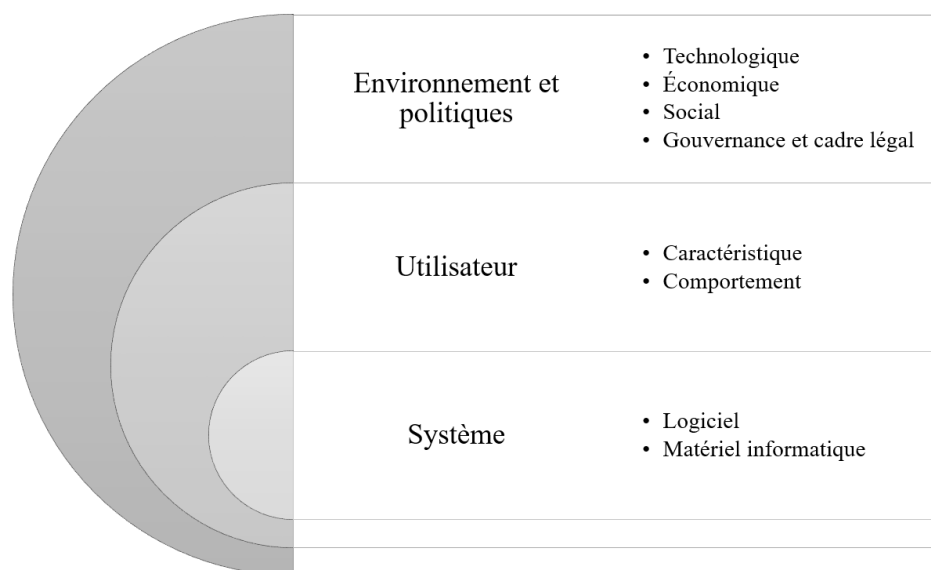


Figure 3.1 Déterminants des attaques par logiciels malveillants

**Système** Parmi les facteurs étudiés au niveau du système, le système d'exploitation (Carlinet *et al.*, 2008; Leukfeldt, 2015; Leukfeldt et Yar, 2016), le navigateur Web (Leukfeldt, 2015; Leukfeldt et Yar, 2016), le type de serveur, et le type de système de gestion de contenu (Vasek et Moore, 2014; Vasek *et al.*, 2016) installé sur un système ont été identifiés comme étant des facteurs significativement reliés au risque d'attaques par logiciels malveillants. Les principales causes sous-jacentes énoncées sont le nombre d'exploits et de vulnérabilités disponibles (Manes, Casper, 2015; Hoffman, Chris, 2016; Vasek et Moore, 2014; Vasek *et al.*,



2016), ainsi que la popularité du système ou du logiciel (Vasek et Moore, 2014; Vasek *et al.*, 2016; Leukfeldt, 2015; Leukfeldt et Yar, 2016; Kaspersky Lab, 2018). Le type et le nombre d'applications installées sur un système ont aussi été identifiés comme étant des facteurs contributifs (Bilge *et al.*, 2017; Ovelgönne *et al.*, 2017). Notamment, un nombre élevé d'applications installées serait associé à un risque d'infections plus élevés.

D'autres chercheurs se sont quant à eux concentrés sur certaines mesures de protection, soit l'utilisation d'un logiciel de protection (antivirus, pare-feu, etc.) (Maier *et al.*, 2011; Wilsem, 2013; Leukfeldt, 2015; Leukfeldt et Yar, 2016), et la mise à jour des applications installées (Bilge *et al.*, 2017; Leukfeldt, 2015; Leukfeldt et Yar, 2016). Dans les deux cas, les résultats suggèrent que ces facteurs contribuent à diminuer le risque d'infections par logiciels malveillants. Similairement, les travaux de Bilge *et al.* (2017) suggèrent un lien positif entre le nombre de vulnérabilités présentes sur un système, et la probabilité d'infections de ce dernier.

**Utilisateur** Au meilleur de nos connaissances, il existe peu d'études dont l'objectif principal est d'analyser l'impact des facteurs démographiques sur le risque d'attaques par logiciels malveillants (Bossler et Holt, 2009; Ngo et Paternoster, 2011). Cependant, certains chercheurs ont tout de même étudié comment ces facteurs sont corrélés aux attaques par logiciels malveillants, bien que cette analyse n'était pas le sujet premier de leur recherche (Wilsem, 2013; Yen *et al.*, 2014; Kranenbarg *et al.*, 2017). Somme toute, ces travaux de recherche suggèrent que certains facteurs démographiques, tels que le genre et l'âge, pourraient être associés au risque d'attaques par logiciels malveillants. Il n'existe toutefois aucun consensus quant à leur effet respectif, c'est-à-dire si ces derniers doivent être considérés comme des facteurs de risque ou des facteurs de protection. De plus, la grande majorité des travaux se limitent à trouver des corrélats significatifs; très peu explorent les mécanismes plausibles qui sous-tendent les facteurs identifiés et leur interdépendance.

D'autres chercheurs ont analysé la relation entre le comportement usager et le risque d'attaques par logiciels malveillants. À cet effet, la recherche existante semble s'entendre sur le fait que certains comportements, tels qu'un volume élevé d'activités en ligne ou la visite de certaines catégories de sites Web, contribuent à augmenter le risque d'attaques par logiciels malveillants (Maier *et al.*, 2011; Canali *et al.*, 2014; Yen *et al.*, 2014).

**Environnement et politiques** Bien qu'il existe plusieurs rapports et données publiées sur les différences géographiques des attaques par logiciels malveillants (voir Section 2.1.4), peu de travaux se sont portés sur l'explication de ces variations. Dans l'ensemble, les tra-

vaux de recherche précédemment cités (voir Tableau 3.2) indiquent que certains facteurs socio-environnementaux seraient reliés au risque d'attaques par logiciels malveillants. Par exemple, il a été trouvé que les pays dont le développement est plus élevé, tant au niveau économique, éducationnel, et technologique, seraient moins susceptibles aux attaques par logiciels malveillants. D'autres travaux suggèrent un effet limité quant à l'adoption de politiques en sécurité de l'information sur le taux national d'attaques (Png *et al.*, 2008; Van Eeten *et al.*, 2010; Kleiner *et al.*, 2013). Cependant, il n'existe aucun consensus dans la littérature quant à l'identification de ces facteurs, et la direction de leur effet respectif. De plus, la majorité des travaux se limitent à identifier des corrélats et offrent peu de discussion sur la nature sous-jacente des associations statistiques identifiées.

### Travaux de recherche proposés

Tel que vu dans la section précédente, la littérature existante rapporte que plusieurs facteurs (éducation, économie, comportement usager, etc.) sont reliés au risque d'attaques par logiciels malveillants. En s'inspirant des méthodes utilisées en épidémiologie et de la littérature sur les déterminants de la santé, nous souhaitons, similairement, identifier quels sont les déterminants qui sont reliés aux attaques par logiciels malveillants. Plus particulièrement, nos travaux portent sur l'identification de facteurs socio-environnementaux, démographiques, et comportementaux.

**Facteurs socio-environnementaux** *Quels sont les facteurs socio-environnementaux qui sont reliés au taux national d'infections par logiciels malveillants ?*

Cette partie de la thèse vise à : i) décrire les différents patrons nationaux d'agrégation d'attaques par logiciels malveillants, ii) étudier comment l'effet des déterminants au niveau national varie pour différentes populations, et iii) générer des hypothèses étiologiques sur la nature des relations identifiées. À cet effet, la méthodologie préconisée prend son inspiration de l'épidémiologie et de l'étude des populations. Plus spécifiquement, nous avons réalisé une étude écologique multi-pays. Ce type d'étude de nature observationnelle est particulièrement utile pour identifier des facteurs de risque et des facteurs de protection haut niveau au sein des populations. Ces populations peuvent soit être définies par unité temporelle ou spatiale ; l'unité dans notre cas étant le pays. Les détails de cette étude sont présentés au Chapitre 4, et ont fait l'objet d'un article qui a été publié en 2016 à la conférence *Workshop on the Economics of Information Security (WEIS)* (Lalonde Lévesque *et al.*, 2016).

**Facteurs comportementaux** *Quels sont les comportements des usagers qui sont reliés au risque d'attaques par logiciels malveillants ?*

Cette partie de la thèse est basée sur une étude utilisateurs réalisée en 2011-2012 par Lalonde Lévesque *et al.* (2013). Cette étude visait, entre autres, à étudier l'interaction entre les usagers, les solutions antivirus, et les logiciels malveillants. Dans le cadre de la présente thèse, une analyse supplémentaire des données comportementales a été réalisée afin d'identifier les comportements des usagers les plus à risque. Les résultats de cette analyse sont présentés au Chapitre 5, et ont fait l'objet d'un article qui a été publié en 2018 à la revue *ACM Transactions on Security and Privacy (TOPS)* (Lalonde Lévesque *et al.*, 2018).

**Facteurs démographiques** *Quel est l'effet indépendant de l'âge du genre sur le risque d'attaques par logiciels malveillants ?*

Cette partie des travaux vise à : i) identifier les populations d'usagers les plus à risque, ii) étudier comment l'effet des facteurs démographiques varie pour différents types de logiciels malveillants, et iii) générer des hypothèses étiologiques, si applicable, sur la nature des relations identifiées. Pour ce faire, nous avons réalisé une étude cas-témoins. Ce type d'étude est particulièrement utilisé en épidémiologie afin de déterminer les facteurs qui contribuent au risque de développer une maladie. De nature observationnelle, une étude cas-témoin consiste à suivre un groupe d'individus et de comparer sur la base du facteur d'intérêt le groupe qui a développé la maladie (cas) au groupe qui n'a pas développé la maladie (témoins). Dans notre cas, l'âge et le genre sont considérés comme facteurs d'intérêts, et la maladie est le fait d'avoir été exposé à une attaque par logiciel malveillant. Les résultats de cette étude ont été publiés en 2017 à la *British Human Computer Interaction Conference (BHCI)* (Lalonde Lévesque *et al.*, 2017), et sont décrits au Chapitre 6.

### 3.2.2 Efficacité des logiciels antivirus

*Quelle est l'efficacité réelle des logiciels antivirus ?*

La section qui suit vise en un premier temps à introduire le lecteur à l'industrie des tests antivirus, ainsi qu'aux méthodes d'évaluation existantes. En un second temps, nous proposons une nouvelle approche d'évaluation des solutions antivirus basée sur la santé publique. Finalement, nous exposons et mettons en contexte les travaux de recherche proposés dans le cadre de la présente thèse.

## Industrie des tests antivirus

Plusieurs organisations indépendantes se spécialisent dans l'évaluation des logiciels antivirus. Parmi les plus notables, nous retrouvons entre autres AV Test, NSS Labs, AV Comparatives, SE Labs, et Virus Bulletin. Ces différentes organisations offrent une variété de tests allant du test de certification au test comparatif d'antivirus. Alors que certains tests visent à évaluer les fonctionnalités globales des produits antivirus, d'autres se spécialisent sur des aspects précis tels que le taux de détection, le taux de faux positifs, l'utilisabilité, la désinfection, ou encore la performance. Ce dernier type de test vise à évaluer plusieurs aspects du produit antivirus comme l'utilisation de la mémoire, le temps de démarrage à l'ouverture de l'ordinateur, la vitesse d'analyse de fichiers, etc.

Les tests de logiciels antivirus ne sont cependant pas encadrés par des normes ou une réglementation officielle. Toutefois, plusieurs efforts ont été faits en ce sens afin de fournir des lignes directrices. Différentes organisations telles que EICAR (European Institute for Computer Antivirus Research) et AVIEN (Anti-virus Information Exchange Network) ont développé leurs propres codes de conduite en ce qui concerne les tests de logiciels antivirus. En 2008, l'organisation Anti-Malware Testing Standards Organization (AMTSO) a été créée afin d'améliorer la qualité, la pertinence et l'objectivité des tests d'antivirus. Regroupant plusieurs acteurs tant du milieu académique qu'industriel, l'organisation vise entre autres à fournir des principes de base aux testeurs de logiciels antivirus (Anti-Malware Testing Standards Organization, 2008). Parmi ces principes, nous retrouvons notamment que les tests réalisés ne doivent pas mettre en danger le public. En d'autres mots, les testeurs doivent s'assurer de ne pas propager les logiciels malveillants utilisés dans le cadre de leurs tests et de ne pas créer de nouveaux logiciels malveillants. Autre aspect important, AMTSO encourage les testeurs à fournir une description de la méthodologie utilisée afin de garantir une meilleure transparence des tests.

## Méthodes d'évaluation existantes

Les méthodes typiques d'évaluation des produits antivirus sont principalement basées sur des tests automatiques réalisés en laboratoire dans des environnements contrôlés. Par exemple, certains tests, dits statiques, consistent à soumettre aux antivirus un échantillon de fichiers composé de logiciels malveillants et de programmes légitimes. Cependant, compte tenu du fait qu'il n'y a pas d'exécution de fichier, et donc aucun comportement logiciel à observer, ce type de tests ne peut refléter adéquatement la capacité des produits qui utilisent des techniques de détection actives et proactives. À l'opposé des tests statiques, les tests dynamiques consistent soit à exécuter des fichiers ou encore à exposer l'antivirus à des sites Web compromis (Anti-

Malware Testing Standards Organization, 2008). Bien que ces tests soient plus réalistes en principe, ils présentent tout de même plusieurs limitations quant à leur capacité d'évaluer la performance réelle des produits antivirus.

**Échantillon de menaces** Un des principaux problèmes avec les tests réalisés en laboratoire se pose au niveau du choix de l'échantillon des logiciels malveillants (Gordon et Ford, 1996; Harley et Lee, 2008; Kosinar *et al.*, 2010). Souvent, cet échantillon est soit trop petit ou aucunement représentatif de la réalité : il contient des logiciels malveillants “fabriqués” pour les tests, ou encore de vrais logiciels malveillants qui ne représentent plus les tendances observées (Kosinar *et al.*, 2010). Un autre facteur important à prendre en considération est la notion du temps. La création d'une banque de logiciels malveillants est une longue opération alors que de nouveaux types de menaces sont créés tous les jours (Muttik et Vignoles, 2008). L'organisation The WildList Organization International a proposé une collection de virus fournis par la communauté. Cette liste a comme principal avantage d'avoir été validée par des experts, ce qui réduit le risque de faux positifs. Elle contient uniquement des virus qui ont été observés à “l'état sauvage”. De plus, elle diminue le risque de certains biais, comme le biais géographique, puisque tous, indépendamment de leur localisation, peuvent y contribuer (Harley, 2009). Cependant, cette liste n'est pas nécessairement représentative de la totalité des logiciels malveillants. Elle n'est mise à jour que mensuellement, ce qui donne le temps aux compagnies d'antivirus de détecter et d'intégrer ces mêmes virus à leur base de données. Il devient donc presque impossible de réaliser des tests avec un échantillon de logiciels malveillants qui représentent les conditions réelles. Afin de partiellement combler ces limitations, l'organisation AMTSO a créé la Real Time Threat List (RTTL). Cette liste vise à fournir en temps réel des informations sur les différentes menaces informatiques telles qu'observées à l'état sauvage par les vendeurs antivirus à l'échelle mondiale (Zwienenberg *et al.*, 2013). Entre d'autres mots, cette liste permet de réaliser des tests basés sur un échantillon qui est représentatif de l'état actuel de l'écosystème des logiciels malveillants. Cependant, bien que de tels tests soient plus réalistes au niveau des logiciels malveillants, ils ne peuvent refléter les performances réelles des produits antivirus, puisqu'ils font, notamment, abstraction de l'utilisateur. Il devient dès lors nécessaire de répliquer dans les tests certains comportements des usagers, tels que la visite de sites Web, le téléchargement de fichiers, la simulation d'attaques basée sur l'ingénierie sociale, ou encore l'exploitation de vulnérabilités.

**Interaction de l'utilisateur** Bien qu'il puisse être difficile de prendre en compte l'interaction de l'utilisateur lors de tests réalisés en laboratoire, certaines organisations tentent d'adresser ce problème en utilisant des applications qui simulent la souris, le clavier et l'interaction avec

de vrais programmes sur l'ordinateur (Vrabec et Harley, 2010). Toutefois, le problème majeur de ces tests est que chaque utilisateur est différent et qu'il est impossible de réaliser un scénario de tests par défaut qui soit suffisamment représentatif de la diversité des comportements usager (Kosinar *et al.*, 2010). Vrabec et Harley (2010) et Muttik et Vignoles (2008) ont proposé comme alternative de créer différents scénarios de tests adaptés selon certains profils utilisateurs. À titre d'exemples, un utilisateur ayant un profil d'internaute peut être simulé par un script qui visite plusieurs sites Web alors qu'un utilisateur ayant un profil de joueur peut visiter différents sites de jeux en ligne et télécharger des jeux. Non seulement l'utilisateur peut être simulé, mais le type de tests peut aussi être adapté selon le profil. Les tests réalisés pour un utilisateur présentant un profil de joueur devraient être orientés vers la latence du réseau ou encore la dégradation de l'affichage par seconde, alors que les tests pour un utilisateur de type travailleur devraient, par exemple, mettre l'emphasis sur le téléchargement de fichiers à partir d'un serveur ou encore l'édition de fichiers vidéo et audio. Le fait d'évaluer les antivirus sous différents profils devrait permettre d'exposer les forces et les avantages de chaque produit dans un contexte se rapprochant d'une utilisation réelle et ce, selon le profil de l'utilisateur. Une première expérience a été réalisée en ce sens en 2013 par PC Security Labs (PC Security Labs, 2013). Regroupant au total sept profils d'utilisateurs, l'expérience a tenté d'évaluer les performances de différentes solutions antivirus en prenant en compte les besoins spécifiques de chaque type d'utilisateur, tout en simulant leurs comportements par des scripts. De par ses résultats, cette expérience a démontré que les performances d'un produit antivirus peuvent varier en fonction du profil de l'utilisateur.

### **Efficacité réelle des antivirus**

Tel que vu dans la section précédente, les méthodes courantes d'évaluation d'antivirus sont basées sur des tests automatisés effectués dans des environnements contrôlés. Bien que ces tests permettent d'évaluer l'efficacité (*efficacy* en anglais) des produits antivirus sous des scénarios spécifiques, ils ne mesurent pas l'efficacité sur le terrain, ou efficacité réelle (*effectiveness* en anglais), des produits déployés sur des machines exploitées par de vrais utilisateurs.

À l'instar des nouveaux médicaments ou des nouvelles interventions médicales qui sont étudiés d'abord en laboratoire et plus tard sur le terrain, les tests antivirus pourraient adopter une approche clinique similaire. Dans le développement de vaccins, par exemple, des études d'efficacité sont utilisées pour évaluer la performance de ce dernier dans des conditions cliniques optimales. Une fois que le vaccin s'est révélé efficace, des études d'efficacité sur le terrain, également appelées études d'efficacité réelle, sont utilisées pour mesurer la protection vaccinale directe et indirecte dans des conditions réelles. Puisque ces conditions sont souvent

sous-optimales par rapport aux conditions cliniques, la protection du vaccin est souvent plus faible que dans les études d'efficacité. De plus, ce type d'études permet d'évaluer comment la protection vaccinale réelle est affectée par des facteurs externes, tels que les facteurs viraux, les facteurs de l'hôte, le stockage, l'administration, la disponibilité et fabrication du vaccin, etc.

Similairement, les produits antivirus pourraient d'abord être évalués dans des conditions contrôlées. En outre, des études d'efficacité réelle pourraient être menées en complément des évaluations en laboratoire. Dans de telles études, les produits antivirus seraient évalués au fil du temps sur des systèmes déployés utilisés dans des conditions réelles. Une telle approche pourrait aider à mieux comprendre comment les produits fonctionnent en utilisation réelle et comment les facteurs externes, tels que l'environnement, la configuration du système et le comportement de l'utilisateur, affectent la performance des antivirus. Les études sur le terrain des produits antivirus pourraient également fournir des informations cruciales aux compagnies d'antivirus sur les aspects du produit (interface utilisateur, détection, mises à jour des fichiers de signatures, etc.) qui pourraient être améliorés.

Une première méthode d'évaluation sur le terrain consiste à mener des essais cliniques, comme proposé en 2009 par Somayaji *et al.* (2009). Dans un contexte de sécurité informatique, cette méthode implique de réaliser des études utilisateurs afin d'évaluer l'efficacité d'un antivirus à protéger un système contre différents logiciels malveillants. En s'inspirant de cette méthode, Lalonde Lévesque *et al.* (2013) ont réalisé une première étude de terrain visant à évaluer un produit antivirus en incluant des usagers réels. Cette étude, qui est une première en son genre, a permis de démontrer le potentiel et la faisabilité d'une telle approche. Bien que cette étude se soit limitée à un seul produit antivirus, la méthodologie peut également être adaptée pour conduire des essais cliniques comparatifs de produits antivirus (Lalonde Lévesque *et al.*, 2012b). Une autre méthode potentielle consiste à mener des études observationnelles. Contrairement aux études expérimentales, telles que les essais cliniques, les produits antivirus ne sont pas installés sur les systèmes. À l'inverse, les systèmes sont surveillés avec leur protection actuelle sans aucune intervention. Par exemple, Blackbird et Pfeifer (2013a) a utilisé des données du Malicious Software Removal Tool (MSRT) provenant de millions de systèmes pour évaluer l'impact de l'état de la protection des produits antivirus sur les taux d'infection.

## Travaux de recherche proposés

Tel que mentionné dans la revue des méthodes d'évaluation des solutions antivirus, la grande majorité des tests actuels sont réalisés dans des environnements contrôlés. Ces tests sont

conçus afin à mesurer l'effet direct des logiciels antivirus, que ce soit dans un contexte individuel ou comparatif. Dans cette optique, les travaux de la présente thèse visent en un premier temps à évaluer l'effet indirect, si effet indirect il y a, des solutions antivirus. En un second temps, nous souhaitons réaliser un test comparatif en conditions réelles afin de mesurer l'effet de l'environnement sur la performance des solutions antivirus.

### **Évaluation agrégée** *Quel est l'état de santé de l'écosystème des logiciels antivirus ?*

Cette évaluation agrégée des solutions antivirus a comme objectifs de : i) mieux comprendre la condition globale de l'écosystème des solutions antivirus, ii) identifier des problématiques qui ne pourraient être étudiées avec les méthodes de tests actuelles, et iii) étudier l'effet agrégé, si effet agrégé il y a, des solutions antivirus. À cet effet, notre recherche est fortement inspirée du concept de santé des écosystèmes. Cette approche, initialement utilisée dans un contexte d'écosystèmes naturels, revient à étudier la condition globale d'un écosystème. En définissant des indicateurs appropriés, il est alors possible de suivre l'activité d'un écosystème ou encore d'en évaluer et prédire les changements (Bertollo, 1998). Dans le cas du présent travail, nous avons considéré la santé d'un écosystème de logiciels antivirus comme étant sa performance agrégée à protéger les ordinateurs contre les attaques par logiciels malveillants. Nous avons développé des indicateurs reliés à l'activité, la diversité et la stabilité de l'écosystème des solutions antivirus. Les résultats de cette évaluation ont été publiés en 2015 à la conférence *International conference on Malicious and Unwanted Software : The Americas (MALWARE)* (Lalonde Lévesque *et al.*, 2015). L'article en question est présenté au Chapitre 7.

### **Évaluation comparative** *Quel est l'impact de l'environnement sur la performance des logiciels antivirus ?*

L'analyse comparative des solutions antivirus a comme principaux objectifs de : i) développer une méthodologie pour évaluer en conditions réelles l'efficacité des solutions antivirus, et ii) déterminer dans quelle mesure l'environnement, tels que le contexte socio-économique, l'utilisateur, et le type de logiciels malveillants, ont un impact sur la performance réelle des solutions antivirus. Par le fait même, nous sommes à même d'identifier des populations de systèmes pour lesquelles les solutions antivirus sont moins efficaces. Nous avons choisi de réaliser une étude observationnelle de type cohortes, aussi appelée étude longitudinale. Ce type d'étude est souvent utilisé en médecine, écologie, psychologie, et science sociale pour déterminer s'il y existe une association entre l'exposition à un facteur et une maladie. Il permet de suivre les changements sur une longue période de temps chez un groupe –cohorte– exposé à un facteur d'intérêt et un groupe similaire qui ne l'est pas. À la fin de l'étude, les



deux groupes sont comparés sur la base de leur taux respectif d'incidence d'une maladie afin de vérifier si le développement de cette dernière est relié au facteur d'intérêt. Dans notre cas, le facteur d'intérêt est le fait d'être protégé par une solution antivirus, et la maladie consiste à être infecté par un logiciel malveillant. Les détails de cette analyse sont présentés au Chapitre 8, et ont été publiés en 2016 à la conférence *International Virus Bulletin Conference* (Lalonde Lévesque *et al.*, 2016a).

### 3.2.3 Source de données

Nous présentons dans la section qui suit les quatre principales sources de données utilisées dans l'application de notre modèle de prévention. Les trois premières sont de nature propriétaire et proviennent de la corporation Microsoft. Quant à la dernière, elle provient d'une étude utilisateurs réalisée par Lalonde Lévesque *et al.* (2013).

#### MSRT

Les données reliées à l'état des ordinateurs ainsi qu'aux infections par logiciels malveillants sont collectées par l'outil MSRT. Cet outil offert gratuitement par Microsoft permet d'analyser les ordinateurs afin de détecter et supprimer certaines familles de logiciels malveillants présentes sous l'environnement Windows. Par conséquent, les familles de logiciels malveillants couvertes par MSRT représentent seulement une partie de l'ensemble des logiciels malveillants présents sous Windows. MSRT est installé et exécuté mensuellement par *Windows Updates* sur plus d'un milliard d'ordinateurs Windows en plus d'être disponible sur demande. Lors de son exécution, MSRT collecte des informations générales sur l'état de l'ordinateur (localisation, système d'exploitation, navigateur Web par défaut, etc.). MSRT fait de plus appel à l'interface de programmation applicative du Windows Security Center (WSC) afin de collecter des informations sur l'état de sécurité de l'ordinateur, tels que le produit antivirus installé et le statut de ce dernier (arrêté, à jour, expiré, etc.).

#### Microsoft Windows Defender

Notre seconde source principale de données provient de Microsoft Windows Defender ; la solution antivirus offerte gratuitement par Microsoft. Alors que MSRT est uniquement exécuté mensuellement ou sur demande, Microsoft Windows Defender offre une protection temps réelle contre les logiciels malveillants. Similairement à MSRT, Microsoft Windows Defender collecte aussi des informations générales sur l'état de l'ordinateur, tel que la localisation, le système d'exploitation et le navigateur Web par défaut. Les données télémétriques de

l'antivirus permettent aussi d'obtenir des informations sur les logiciels malveillants qui ont été bloqués par Microsoft Windows Defender. Ces informations contiennent, entre autres, le nom du fichier détecté, la famille de logiciels malveillants, et le type de logiciels malveillants (cheval de Troie, virus, ver, etc.).

### **Microsoft account**

Les données collectées par MSRT et Microsoft Windows Defender peuvent, dans certains cas, être couplées avec des informations provenant de Microsoft Account ; un système d'authentification unique permettant de se connecter à plusieurs services offerts par Microsoft (par exemple, Outlook, Skype, OneDrive). Par l'intermédiaire de ce système, il est possible d'obtenir des données démographiques sur les usagers de Windows 10, tel que le genre et le groupe d'âge associé.

### **Étude utilisateurs**

Ce jeu de données provient d'une étude utilisateurs réalisée par l'auteure de la présente thèse (Lalonde Lévesque *et al.*, 2013). L'expérience, basée sur le concept des études cliniques, a impliqué la participation de 50 usagers durant quatre mois. Au cours de cette période, des données ont été collectées sur l'état des ordinateurs, ainsi que sur les utilisateurs. Ces données incluent, entre autres, la liste des applications installées, le volume et le type de sites Web visités, les caractéristiques de l'utilisateur, ainsi que l'occurrence et le type d'attaques par logiciels malveillants rencontrées par les utilisateurs.

## CHAPITRE 4 ARTICLE 1 : NATIONAL-LEVEL RISK ASSESSMENT : A MULTI-COUNTRY STUDY OF MALWARE INFECTIONS

*Published in the Proceedings of the 15th Workshop on Information Security (WEIS) 2016.*

**Authors** Fanny Lalonde Lévesque<sup>1</sup>, José M. Fernandez<sup>1</sup>, Anil Somayaji<sup>2</sup>, Dennis Batchelder<sup>3</sup>

<sup>4</sup>

**Institutions** École Polytechnique de Montréal<sup>1</sup>, Carleton University<sup>2</sup>, AppEsteem<sup>3</sup>, Microsoft Corporation<sup>4</sup>

**Abstract** The security of computers is a function of both their inherent vulnerability and the environment in which they operate. Much as with the public health of human populations, the “public health” of computer populations can be studied in terms of what factors influence their security. Using data collected from Microsoft Windows Malicious Software Removal Tool (MSRT) running on more than one billion machines, we conduct a multi-country analysis of malware infections and measures of economic development, educational achievement, Internet infrastructure, and cybersecurity preparedness. We find that while increases in these factors is often correlated with reduced infection rates, their significance and magnitude vary considerably. In contrast to past work, these variations suggest that policy interventions, such as efforts to increase the quality of home Internet connections, are likely to decrease infection rates in only some circumstances.

**Keywords** Risk factor, Malware, Ecological study, Cybersecurity, Population health, Public policy

### 4.1 Introduction

The susceptibility of computers to malware infections is known to be affected by technological factors (e.g. their hardware, operating system, and applications) (Lalonde Lévesque *et al.*, 2013; Maier *et al.*, 2011; Carlinet *et al.*, 2008) and human factors (e.g. computer expertise, risk aversion) (Lalonde Lévesque *et al.*, 2013; Onarlioglu *et al.*, 2012; Sheng *et al.*, 2010). With human health, however, we know that factors such as economic development, geographic location, and the aggregate health choices all influence the health of individuals. For example,

while individuals can take steps, such as using mosquito nets and insect repellent, to avoid catching malaria, the biggest factor influencing whether you may get malaria is simply where you happen to live. If the mosquitoes in your area happen not to carry malaria then you are safer from it—even if you take no other protective steps. Similarly, if authorities in the area you live in take steps to reduce the number of mosquitoes, your risk of malarial infection also goes down, all without any changes in individual susceptibility or individual behavior. Our question here, then, is can we identify analogous factors that could be changed through national policies, such as the prevalence of mosquitoes or vaccination rates, that would improve the security level of entire computer populations?

For instance, while it may be intuitive that wealthier nations perform better in cybersecurity, or that nations with higher Internet connectivity are more susceptible to cybercrime, it is essential to validate those hypotheses and understand their causes. Many studies, mostly from antimalware vendors, security experts, or networking providers, report on geographical patterns and trends in malware infections without investigating the factors behind those variations. So far, only few studies have examined how national factors (e.g. income, education, Internet penetration) correlate with cross-country differences in malware infections (Garg *et al.*, 2013; Mezzour *et al.*, 2015; Subrahmanian *et al.*, 2016; Burt *et al.*, 2014). However, there is no overwhelming consensus in the literature on which factors are the best predictors of malware infections at the national-level. Moreover, those studies offer little or no discussion on potential underlying causes for their findings. Consequently, lack of consensus and scarcity of evidence represent a serious challenge for cybersecurity policy making. In order to support good, evidence-based policy making, we need to conduct empirical studies on large and representative populations of computer systems that will provide understanding of the causes of malware infections.

Fortunately observations of large computer populations is now feasible due to telemetry systems embedded into commonly-used security software. While these systems were originally developed for quality assurance, they can also be used to study the patterns associated with malware infection and other security violations. Security software telemetry data thus allows us to adopt a *population health* approach. Formally, population health refers to “the health outcomes of a group of individuals, including the distribution of such outcomes within the group” (Kindig et Stoddart, 2003)—the population, in our setting, being computer systems. Similarly, a large body of work has also looked at how *public health* may serve as a model for cybersecurity (Rowe *et al.*, 2012a; Rice *et al.*, 2010; Rowe *et al.*, 2013; Mulligan et Schneider, 2011; Sullivan *et al.*, 2012; Charney, 2010). Much as with health, epidemiological techniques can then be applied to security to investigate factors and conditions that affect the health status of computer systems in order to develop cybersecurity policies and strategies that

reduce the risk of security compromise.

There have been a few previous epidemiological studies that used security telemetry data to identify risk factors related to malware infection (Thonnard *et al.*, 2015; Yen *et al.*, 2014). While these past studies have identified technical and behavioral factors related to individuals and organizations, they were not designed to identify risk-modifying factors at the national-level. Moreover, interventions focused on individuals or organizations are unlikely to succeed if the environmental condition in which they are delivered are not supportive. Therefore, there is a need to understand the multi-level risk factors leading to malware infection, including both *proximal*, *intermediate* and *distal* factors, as the latter two are often determinants of the risk factors. While proximal factors act directly or almost directly on the cause of infection, distal factors are further back in the causal chain and act via a number of intermediate factors. As both ecological along with individual and behavioral determinants play an important roles in the development and prevention of malware infection, it becomes important to conceptualize the problem within multiple levels of influence.

Commonly used within population health research, *ecological studies* can be designed to identify risk factors at higher levels. In such studies, populations are defined by temporal (tracking a population over an extended period or time) or spacial (comparing populations in different geographic locations) units and compared on their prevalence or incidence of disease. This type of observational study is particularly useful for generating and testing hypotheses on potential risk factors, whether distal, intermediate, or proximal. From there, other epidemiological or laboratory approaches can be used to test the causality, if any.

In this paper we report on a multi-country ecological study of risk factors related to malware infections. Country infection rate is computed using large-scale telemetry data from millions of systems running Microsoft Malicious Software Removal Tool (MSRT), a malware cleaner utility that scans Windows systems for infections of specific malicious software. We investigate association of factors related to economics, education, technology, and cybersecurity on malware infection rate by country. We develop regression models for the prevalence of malware infection to identify and quantify the relative importance of those risk factors and how their effects vary between countries with different socio-economic status. In summary, our main contributions are:

1. We present a multi-country ecological study of malware infection risk factors, based on a large sample of unprotected hosts (100 million).
2. We investigate how malware infections at the national-level correlate with factors related to economy, education, technology and cybersecurity, including some previously unstudied factors like antivirus penetration, Global cybersecurity index, etc.

3. We develop a regression model for the prevalence of malware infection that identifies and quantifies the relative importance of those factors and how their effect vary between countries with different socio-economic status.
4. We identify potential risk-modifying factors that can be influenced by cybersecurity policies.

The remainder of the paper is organised as follows. Section 6.2 presents a review of previous studies and Section 6.3 describes the study in terms of data collection and analysis. In Section 6.4 we present the results in terms of national-level risk factors for malware infection. We discuss our observations and study limitations in Section 6.5. We conclude and discuss potential implications of our findings in Section 6.7.

## 4.2 Previous studies

We present a short review of past work focusing on the link between national factors and malicious attacks at a cross-country level. In what follows we distinguish our study with prior research in terms of datasets, study design, and analysis methodology.

Some researchers have focused on the impact of national cybersecurity policies on malicious attacks at the country level. Ivan *et al.* Png *et al.* (2008) adapted the event study methodology from research in financial economics to study the impact of government enforcement and economic opportunities on information security attacks in the US. They found limited evidence that domestic enforcement deters attacks within the country. Microsoft also sought to understand whether certain policies can measurably reduce cyberrisks at the national level Kleiner *et al.* (2013). They conducted a descriptive analysis and found that countries adopting or implementing certain policies, like the London Action Plan (LAP) or the Europe Convention on Cybercrime (ECC), may contribute to reduce the risk of malware infection. Overall, those studies more or less all found that national cybersecurity policies may contribute to reduce the risk of malicious attacks.

Garg *et al.* Garg *et al.* (2013) performed a cross-country empirical analysis to investigate how macroeconomic factors grounded in traditional theories of crime offline relate to the rate of machines acting as spambots. Factors related to the availability of machines, guardianship, economic deprivation, legal framework and governance were investigated. Results suggested that higher Internet adoption, measured by the total number of fixed broadband Internet subscribers, is related with a higher percentage of spambots while countries with higher secure Internet servers (per million people) were associated with a lower percentage of spambots.

In another study, Microsoft did a cross-country analysis of different social and economic

policy indicators to predict the rate of malware infections within countries Burt *et al.* (2014). They used 2013 data from MSRT and defined the infection rates as the number of computers cleaned for every 1,000 executions of MSRT. Their predictive model identified 11 factors related to digital access, institutional stability and economic development. Countries with above-average development across those areas were expected to see greater improvement in cybersecurity. Although the authors included in their study a broad set of national factors, their statistical analysis was predictive, and not explanatory. That is, the purpose of their statistical model was to predict the rate of malware infections, which is different from causal explanation. In opposition, our study used explanatory modeling for testing potential causal factors behind international differences in malware infections.

Only few explanatory research have investigated the effects of multiple factors in terms of economics, technology, and cybersecurity on malware prevalence at the national level. Mezzour *et al.* (2015) performed an empirical study to understand how the average malware encounters rate of home users vary internationally. Using 2009-2011 telemetry data from the Symantec's Worldwide Intelligence Network Environment (WINE) (Dumitras et Shou, 2011), the authors empirically test the validity of specific factors related to computing and monetary resources (i.e. GDP per capita, Internet bandwidth, ICT development index), cybersecurity expertise (as measured by cybersecurity research and the existence of cybersecurity institutions), international relations, computer piracy and web browsing. They found that high piracy rates was the main factor associated with high malware encounters especially in countries with low computing resources.

Subrahmanian *et al.* (2016) also leveraged the WINE telemetry data from host machines protected by Symantec's antivirus products. They computed the average number of infection attempts per host of a given country as a proxy of its level of cyber-vulnerability. In an attempt to explain international differences in attack frequencies, they performed a multivariate analysis including macroeconomic factors (per capita GDP, Internet penetration, software piracy) and host-based features aggregated at the country-level (total number of binaries installed, fraction of downloaded binaries, of unsigned binaries, and of low-prevalence binaries). Overall, they found per capita GDP and fraction of downloaded binaries to be significant predictors; countries with low economic wealth (as measured by per capita GDP) and high fraction of downloaded binaries were more vulnerable. In contrast to Mezzour *et al.* (2015), they found software piracy to be non-significant, suggesting that its effect may be more a function of other variables, such as per capita GDP, than a direct cause of cyber-vulnerability. In comparison to Mezzour *et al.* (2015) and Subrahmanian *et al.* (2016), our research is distinct in three important ways. First, the sample population is different; they studied protected host machines (from Symantec) of home users, and we focus on unpro-

tected host machines including both home and corporate users. Second, their dependent variable was the average malware encountered by computer in each country, while we are interested in countries' malware infection rate. Three, our work accounts for a broad set of national factors; neither Mezzour *et al.* (2015) or Subrahmanian *et al.* (2016) investigated factors related to both economy, education, technology, and cybersecurity readiness.

The key way our research differs from past work is in how we designed our study and performed our analysis. While past studies have focused on the identification of national factors (Garg *et al.*, 2013; Mezzour *et al.*, 2015; Subrahmanian *et al.*, 2016) or the development of a predictive model Burt *et al.* (2014), our research goes beyond previous work as we also quantify the relative importance of the studied factors. We also evaluate how the direction and magnitude of those factors vary between countries with different socio-economic development levels, while all research previously cited is limited to a global analysis. Moreover, most of the papers cited above offer little discussion of how their results in terms of national factors should be interpreted, for instance, whether they should be seen as direct or indirect effect or whether they are confounded by other factors. Finally, compared to previous work, our study is grounded in traditional epidemiological techniques.

### 4.3 Study design and methods

A multi-country ecological study was conducted in order to identify which national factors are the best predictors of malware prevalence across countries. This type of observational study was selected as it is often used to identify factors on health when the outcome is averaged for the population in geographical or temporal units. The main advantage is that it allows to study variables that cannot be measured at the individual level or that may have a different effect at the individual and population level. Such variables, called *ecological factors*, can be classified as *aggregate*, *environmental* or *global variables*, depending on what they measure. Aggregate factors are data based on individuals aggregated at the population level. Environmental factors relate to the characteristics of the environment in which people live. Although they are measured at the population level, they can also be measured at the individual level. Global factors are variables computed from groups, organizations, or places for which there is no analogue at the individual level. While ecological studies are convenient to test multiple hypothesis at the same time, special care should be taken to select the appropriate sample size and sampling method, limit potential bias and effect of chance, and control for potential *confounding variables* —undesirable factors that may influence the results and threaten the internal validity of the study.

The data was collected by Microsoft Malicious Software Removal Tool (MSRT), a malware



cleaner utility that scans computers for infections of specific, prevalent malicious software and helps remove these infections (Microsoft Corporation, 2016). MSRT is delivered and runs every month on more than one billion machines through Windows Update as well as being available as a separate download from Microsoft. Upon its execution, MSRT also calls the Windows Security Center (WSC) API to collect information about the protection state of computers, such as the antivirus (AV) actively protecting the machine and its signature status. Such information is then reported by MSRT to Microsoft for a random sample of 10% of the machines. The data used in this analysis was —monthly— collected from June to September 2014 on computers running Windows XP, Vista, 7, 8 and 8.1., which represents 100+ million computers.

### 4.3.1 Data collection

The dependent variable under consideration is the rate of malware infection by country for unprotected computer systems, which represent approximately 10% of the 100+ million computers. The rate of malware infections was computed based on the proportion of unprotected systems that reported at least one infection over the 4 months. Systems were considered unprotected if they had no AV product enabled on their machine. We chose to focus on unprotected systems so as to avoid the bias other AV software would potentially introduce into rates of infection. Moreover, it allows us to focus on malware infections, rather than malware encounters. As far as we know, this is the largest study on malware infections on unprotected systems. Furthermore, Internet Protocol (IP) geolocation was used to identify the country associated with each user report.

Independent variables were selected based on two criteria, i.e. 1) they were plausible risk factors, and 2) they constituted factors that might be possibly reduced by intervention at the country level. We selected 15 factors (see Table 4.1) to cover the socio-economic and technological reality of countries, as well as their level of cybersecurity. A detailed description of factors considered in the current study is presented in the following text and in Appendix A.

**Economic performance** We used the Gross Domestic Product (GDP) and the Gross Domestic Product per capita by purchasing power parity (GDP-PPP) from the World Bank (WB) (The World Bank, 2017), as indicators of the economy. While GDP measures the wealth within a country, GDP-PPP embeds a measurement of income inequality across countries. The direction in which those variables may play is difficult to predict. On the one hand, the economy of countries could influence their resources and opportunities to make choices that could protect their population (Garg *et al.*, 2013). On the other hand, higher monetary

Table 4.1 Country-level factors

Model	Description	Year	Source
Economy	GDP	2013	WB
	GDP-PPP	2013	WB
Education	Mean years schooling	2013	UNDP
Technology	%Households with computer	2013	ITU-D
	%Households with Internet	2013	ITU-D
	Fixed Internet subscriptions (per 100 people)	2012	ITU-D
	Fixed broadband subscriptions (per 100 people)	2013	ITU-D
	Fixed (wired) broadband speed	2013	ITU-D
	%Fixed broadband subscriptions (256kbit/s - 2Mbit/s)	2012	ITU-D
	%Fixed broadband subscriptions (2Mbit/s - 10Mbit/s)	2012	ITU-D
	%Fixed broadband subscriptions (above 10Mbit/s)	2012	ITU-D
	International Internet bandwidth (per million people)	2013	ITU-D
Cybersecurity	Secure Internet servers (per million people)	2013	WB
	%Protected	2014	MSRT
	Global cybersecurity index	2014	ITU-D

resources may cause an increase in malicious attacks, as many malware have a monetary goal (Mezzour *et al.*, 2015). Those factors are global variables and were considered control variables in the analysis as they may be markers for variables we cannot measure nor control.

**Education** As a measure of the level of education, we used the mean years of schooling (MYS) from the United Nations Development Programme (UNDP) (United Nations Development Programme, 2015b), which represents the average number of years of education received by adults aged 25 and older. This variable may account for a possible direct effect of aggregate education and for indirect effects, that are not captured by other factors, such as user behaviour, information technology (IT) literacy, and cybersecurity awareness. We expect that education will be negatively associated with infection rates, as it may affect, among others, users' ability to understand IT information, and follow guidelines for their online safety. This factor is an aggregate variable and was considered as a control variable for the purpose of the analysis.

**Technology** Technological factors were selected from the International Telecommunication Union Development Sector (ITU-D) (International Telecommunication Union, 2015) to capture both the quantity and quality of information and communications technology (ICT). The quantity was evaluated in terms of the percentage of households with a computer, the percentage of households with Internet access, the number of fixed Internet subscriptions (per 100 people) and the number of fixed broadband subscriptions (per 100 people). While factors

related to technology quantity are aggregate variables, we are interested in their environmental effect. For example, countries with large population of computers and Internet users could be more subject to malicious attacks as they may have more potentially vulnerable machines (Garg *et al.*, 2013; Mezzour *et al.*, 2015).

The indicators for the quality were selected to measure both the broadband speed and the bandwidth. For broadband speed, we used the fixed (wired) broadband speed (Mbit/s) (FBS), which refers to the advertised maximum theoretical download speed; it does not refer to the actual speed delivered. We also used the percentage of fixed broadband subscriptions for different speed categories: advertised downstream speed between 256 kbit/s and less than 2Mbit/s (%FB(256-2)), between 2 Mbit/s and less than 10Mbit/s (%FB(2-10)), and greater than or equal to 10 Mbit/s (%FB(10+)). To evaluate the bandwidth, we used the international Internet bandwidth (IIB), which refers to the total used capacity of international Internet bandwidth, in megabits per second. It measures the sum of used capacity of all Internet exchanges offering international bandwidth. We divided the IIB by the country's total population and multiplied by 1 million to obtain the bandwidth by 1 million inhabitants. The direction of the association between technology quality and malware infection rates is difficult to say in advance. As one could argue that technology quality may affect users' ability to protect themselves (i.e. having an up-to-date system or performing signature updates for anti-malware products), it may also contribute to increased cybercrime (remote attacks, spam distribution, software piracy, etc). The factors related to the technology quality were considered aggregate variables.

**Cybersecurity** Factors were selected to capture both private and individual investment in security, and national cybersecurity development. Similar to Garg *et al.* (2013), we used the number of secure Internet servers (per million people) from the WB as a proxy for private investments in security. For the investment at the individual level, we used the percentage of users protected by antivirus product. This last factor was obtained from MSRT and is based on the percentage of systems that have at least one antivirus product actively running with up-to-date signatures during the 4-month period. To evaluate the level of cybersecurity development of countries, we used the Global cybersecurity index (GCI) (Menting, 2014). This index was developed by an ITU-ABI research joint project to rank the cybersecurity capabilities of nation states within five categories: legal measures, technical measures, organizational measures, capacity building and cooperation. We expect all variables to have a negative association with the rate of malware infections. Factors related to cybersecurity at the private and individual level were both considered aggregate variables, while the GCI was a global variable.

### 4.3.2 Statistical methods

The goals of the statistical analysis were 1) to estimate variations in the prevalence of malware infection across countries, 2) to quantify the relative importance of national factors in this variation, and 3) to study the relationship between specific factors and malware infection rates.

While our data set is large overall, for some countries our sampled population is too small to allow for proper analysis. In order to determine the minimum representative sample size for each country, we performed a power analysis to identify the minimum number of computer system reports required. We used a two-tailed one proportion Chi-Square test with a desired power of 90% and a level of significance of 1%. The minimum sample size computed was 37 149 system reports by country, which was rounded to 38 000. We then excluded all countries that had less than 38 000 reports over the 4 months, reducing our sample from 241 to 187 countries.

We implemented a general linear regression model—a specific generalized linear model. First, we ensured that the relationship between the dependent and independent variables was linear, and applied when required a log transformation to the independent variables in order to meet the linearity assumption. All factors were log transformed, except mean years of schooling, %Households with computer, %Households with Internet, %Protected and Global cybersecurity index. Descriptive statistics of the factors before and after the transformation are presented in Table 4.2 and Table 7.5 respectively. The mean allows to measure the central tendency of the data and the standard deviation measures how concentrated the data are around the mean; the more concentrated, the smaller the standard deviation (SD).

In order to identify and assess the unique impact of each factor, we looked for multicollinearity—strong correlations between the independent variables—as it can reduce the amount of information available to evaluate the effect of the factors. The presence of multicollinearity was investigated by computing the variance inflation factor (VIF), which estimates how much the variance of a coefficient is inflated because of linear dependence with other variables. A low value ( $VIF < 5$ ) implies that the variable is uncorrelated with all the other variables (Akinwande *et al.*, 2015; Zainodin *et al.*, 2015). To the opposite, a high value is a sign of multicollinearity. We excluded GDP-PPP ( $VIF > 10$ ) and retained GDP as an indicator of economic performance. Secure Internet servers (per million people) was also found to be highly multicorrelated ( $VIF > 10$ ) with other variables. This factor was excluded, while we retained the Global cybersecurity index and the percentage of users protected by antivirus product as indicators of cybersecurity. All factors related to technology quantity were excluded as they all presented high multicollinearity ( $VIF > 80$ ). The remaining nine factors

Table 4.2 Descriptive statistics (whithout transformation)

Factor	Mean	SD
%Infected	0.22	0.13
GDP	4.82e+11	1.64e+12
GDP-PPP	1.91e+04	2.03e+04
Mean years of schooling	8.29	1.29
%Households with computer	45.46	30.45
%Households with Internet	41.79	30.38
Fixed Internet subscriptions (per 100 people)	14.12	12.67
Fixed broadband subscriptions (per 100 people)	12.93	12.97
Fixed broadband speed	4.48	8.61
%Fixed broadband subscriptions (256kbit/s - 2Mbit/s)	0.35	0.35
%Fixed broadband subscriptions (2Mbit/s - 10Mbit/s)	0.34	0.24
%Fixed broadband subscriptions (above 10Mbit/s)	0.34	0.33
International Internet bandwidth (per million people)	2.63e+05	1.34e+06
Secure Internet servers (per million people)	4.16e+02	9.53e+02
%Protected	0.20	0.09
Global cybersecurity index	0.33	0.22

Table 4.3 Descriptive statistics after log-transformation

Factor	Mean	SD
%Infected	0.22	0.13
GDP*	10.83	0.88
GDP-PPP*	4.03	0.51
Mean years schooling	8.28	2.92
%Households with computer	45.46	30.45
%Households with Internet	41.79	30.38
Fixed Internet subscriptions (per 100 people)*	0.76	0.80
Fixed broadband subscriptions (per 100 people)*	0.60	0.97
Fixed broadband speed*	0.18	0.61
%Fixed broadband subscriptions (256kbit/s - 2Mbit/s)*	-0.87	0.78
%Fixed broadband subscriptions (2Mbit/s - 10Mbit/s)*	5.10	1.22
%Fixed broadband subscriptions (above 10Mbit/s)*	-0.87	0.85
International Internet bandwidth (per million people)*	4.55	0.97
Secure Internet servers (per million people)*	1.62	1.15
%Protected	0.20	0.09
Global cybersecurity index	0.33	0.22

\*Variables have been log-transformed.

all had VIF values under five.

To further evaluate if a linear regression model is appropriate for the data, we performed a graphical analysis of the residuals —the difference between the observed value of the dependent variable and the expected value. The goals of the analysis were to examine if the

residuals 1) have a constant variance, 2) have a mean of 0, and 3) are normally distributed. Results of the residual analysis (see Appendix B) suggested that a linear regression model is adequate. China was also identified as an outlier according to our regression model—it had one of the lowest infection rates (2%), while the regression model predicted a value of 23%. This low infection rate is consistent with recent observations and reports from Microsoft (Microsoft Corporation, 2014b,c). However, research conducted by Microsoft suggested that these low infection rates, as measured by MSRT, may not reflect the threat landscape in China (Rains, 2013). Moreover, as many systems in China use third-party software for update and patch management instead of Windows Updates, those systems are more likely to be fully patched and protected in ways that can’t be measured by MSRT. Based on those potential bias, along with the residual analysis, we decided to exclude China from our regression model, reducing our sample to 186 countries.

## 4.4 Results

The five less infected countries were Aland Islands (1.4%), Japan (3%), Cayman Islands (3.4%), Finland (3.6%) and Liechtenstien (3.7%), while the five most infected countries were Ethiopia (63.8%), Iraq (54.5%), Pakistan (54.2%), Yemen (51.8%) and Sudan (51.2%). As illustrated in Figure 4.1, Africa and South Asia had the highest infection rates while North America and Europe had the lowest.

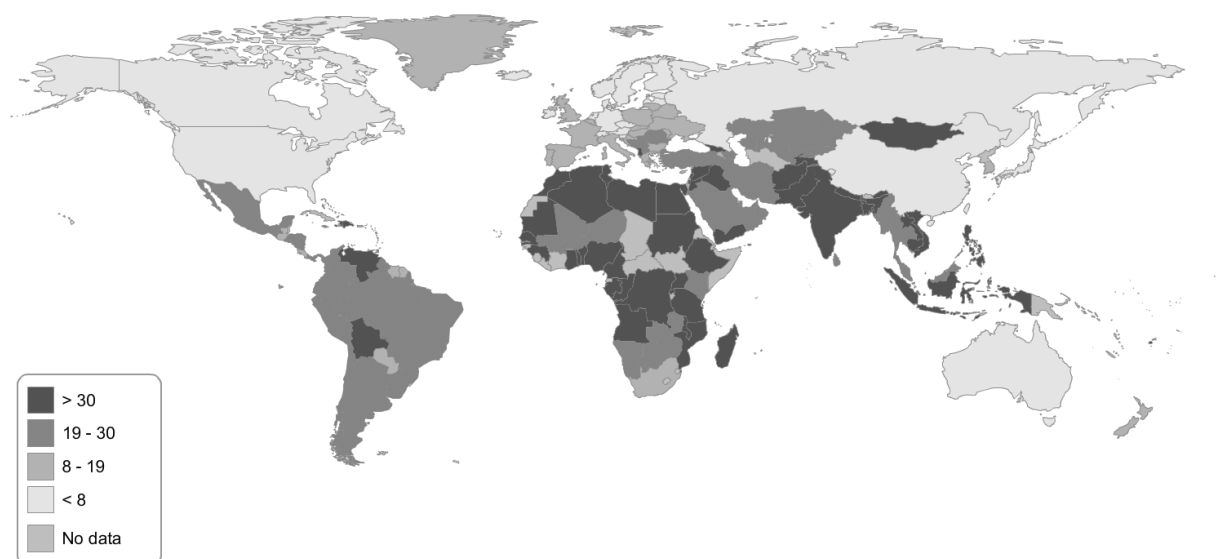


Figure 4.1 Global map of malware infection rates

To better understand the geographical variations, we correlated the infection rate of the

countries with factors measuring economic performance, education, technology and cybersecurity using a general linear regression model. Finally, the same analysis was conducted after categorizing countries by their socio-economic status.

#### 4.4.1 Global model

In order to study how each factor individually relates to the dependent variable, we computed the Pearson correlation coefficients between the infection rate and the nine country-level factors (see Table 4.4). The value  $r$ , the correlation coefficient, represents the strength of the relationship between the variables. The value ranges between -1 and 1, with a value of 0 indicating that there is no linear correlation between the variables. As not all factors were available for the 186 countries, we reported the sample size (N) used for each factor. The p-value was also computed to measure the significance of the results. A low p-value (such as 0.01) means that there is a 1 in 100 chance that we would have obtained the same results if the variables were not correlated. For the purpose of the analysis, we considered a correlation to be significant if the p-value was lower than 0.05.

Table 4.4 shows that all factors are highly significantly ( $p\text{-value} < 0.001$ ) correlated with the infection rate. Except for the variable %Fixed broadband subscriptions (256kbit/s - 2Mbit/s) that has a positive correlation, all other variables were found to have negative association with the infection rate.

Table 4.4 Pearson correlation coefficients between infection rate and country-level factors

Factor	r	N	p-value
GDP-log	-0.37	127	1.86e-05
Mean years schooling	-0.75	145	1.25e-27
Fixed broadband speed-log	-0.57	148	1.03e-13
%FB (256kbit/s - 2Mbit/s)-log	0.53	71	2.03e-06
%FB (2Mbit/s - 10Mbit/s)-log	-0.38	79	4.38e-04
%FB (above 10Mbit/s)-log	-0.72	65	1.21e-11
IIB (per million people)-log	-0.69	147	5.37e-22
%Protected	-0.83	186	0.00e-01
Global Cybersecurity Index	-0.38	154	1.02e-06

Although the Pearson correlation coefficients provide insights on the dependence between two variables, it is very difficult to draw conclusions about the effect of one single factor on the dependent variable. We therefore conducted a multiple general linear regression to estimate the effect of each factor while controlling for the other factors that simultaneously affect the dependent variable. Detailed results of the regression are presented in Table 4.5.

Table 4.5 Global multiple general linear regression results (N=50 countries)

Factor	$\beta$	Std. Error	t-value	p-value
GDP-log	-0.14	0.10	-1.44	0.16
Mean years schooling	-0.31	0.08	-3.61	8.66e-04***
Fixed broadband speed-log	-0.05	0.07	-0.65	0.52
%FB (256kbit/s - 2Mbit/s)-log	-0.13	0.09	-1.47	0.15
%FB (2Mbit/s - 10Mbit/s)-log	0.25	0.10	2.46	1.86e-02*
%FB (above 10Mbit/s)-log	0.03	0.12	0.25	0.80
IIB (per million people)-log	-0.28	0.07	-3.65	7.61e-04***
%Protected	-0.65	0.10	-6.49	1.07e-07***
Global Cybersecurity Index	0.02	0.07	0.31	0.75
R <sup>2</sup> adjusted		0.86		
F-statistic		34.30		
Degree of freedom		9		
Df (residuals)		39		
p-value		7.77e-16		

\*Statistically significant at 0.05 level; \*\*Statistically significant at 0.01 level; \*\*\*Statistically significant at 0.001 level.

For each factor, the standardized regression coefficient  $\beta$  and its associated standard error (Std. Error) were computed. The p-value, which is interpreted as an indicator of the significance of the results, was also computed: a low p-value indicates that the null hypothesis can be rejected with high confidence, and that the variable is relevant in the regression model. We also provided the t-value of each factor, which provides insight on the direction (positive or negative) and magnitude of the effect. The number of countries (N) used for each regression model is also provided. As not all factors were available for the 186 countries, we applied a casewise deletion method, also known as listwise deletion, to handle missing data. With this method, observations that have missing values in at least one factor are removed from the analysis. While such an approach reduces the number of countries, it has the advantage of keeping each studied variable with exactly the same number of observations.

From the regression (see Table 4.5), we can see that the main factors of malware infections are %Protected, international Internet bandwidth, mean years of schooling, and the percentage of fixed broadband subscriptions between 2Mbit/s and 10Mbits/s. As expected, %Protected and mean years of schooling are negatively correlated with the dependent variable. Our results also support that the quality of technology in a country may have an important effect on the rate of malware infections. Bandwidth was found to present a strong negative relationship with the infection rate while broadband speed, as measured by %FB(2-10), presents a weak positive association with the infection rate. Surprisingly, GDP and the GCI were not found to be significant after controlling for the other factors, as opposed to



the results from the Pearson correlations (see Table 4.4). One plausible explanation is that technology quality, education and users' investment in security are channel variables between GDP, the GCI and malware infections. This would imply, for example, that GDP *per se* is not a significant factor for malware infections.

We used a Pareto chart (see Figure 4.2(a)) to visualize the relative importance of the nine country-level factors on the infection rate. The chart displays the absolute value of the effects (as measured by the t-value) and draws a reference line; any factor that extends beyond this line has a statistically significant impact ( $p\text{-value} < 0.05$ ) on the infection rate. The main factor appears to be users' investment in security, as measured by %Protected. Bandwidth and education were found to be equivalent in their effect on the dependent variable, followed by %FB(2-10).

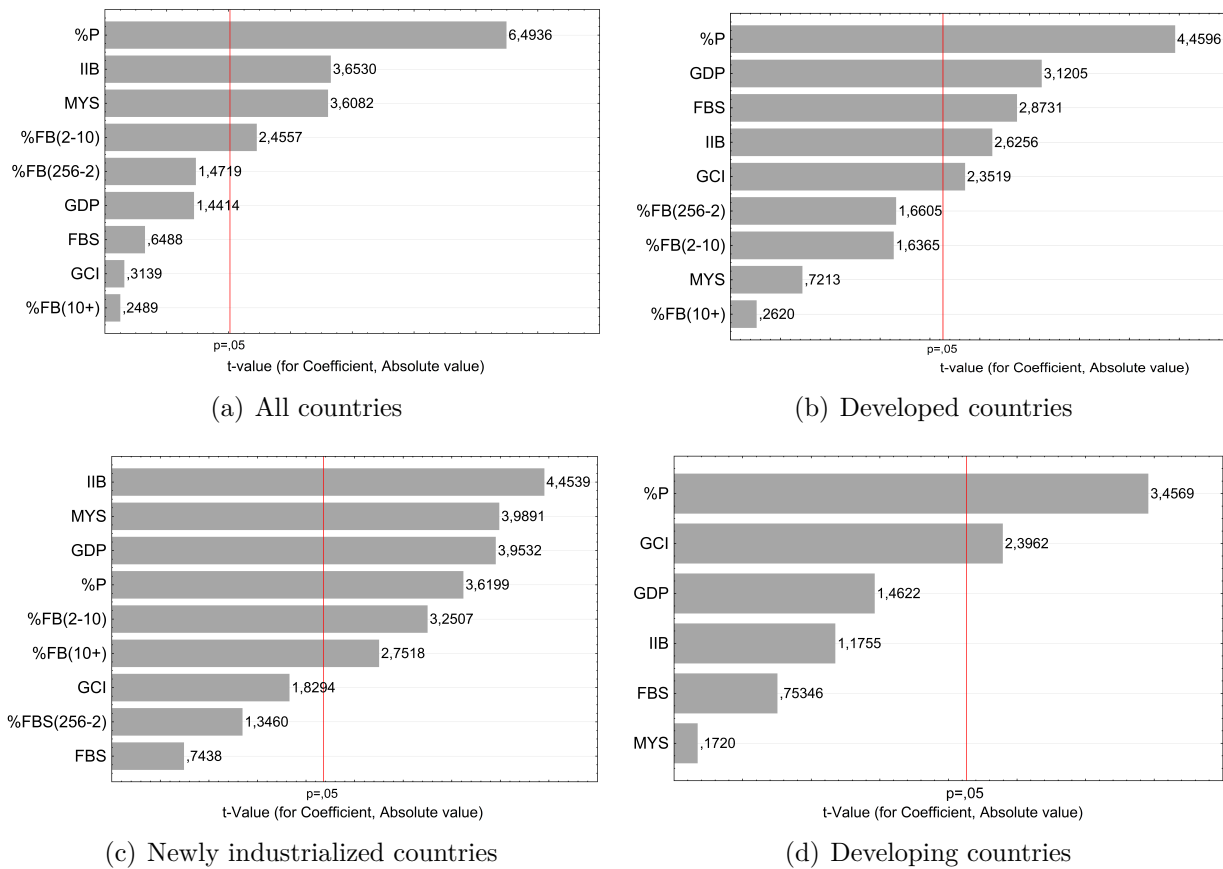


Figure 4.2 Pareto charts by socio-economic status

To evaluate the regression model we used the adjusted  $R^2$ , also known as the coefficient of determination. This number can be interpreted as how well the regression model can explain the variance of the dependent variable. In general, models with values over 80%

are considered strong and models with values over 90% very strong. Overall, our regression model offers a strong prediction ability with an adjusted  $R^2$  of 86%. This indicates that the nine country-level factors selected can explain 86% of the infection rate, a value that is quite high in regard to the literature known to the authors.

#### 4.4.2 Model by socio-economic status

To investigate whether our previous findings apply in countries with different socio-economic development levels, we repeated all analyses after categorizing countries based on their 2013 Human Development Index (HDI) Malik et Jespersen (2013). Overall, 45 countries were considered as developed ( $HDI \geq 0.8$ ), 74 as newly industrialized ( $0.8 > HDI \geq 0.55$ ), and 26 as developing ( $0.55 > HDI$ ), which give us a sample of 145 countries.

As Figure 4.3 illustrates, there is an important variation in the malware infection rates between each category. Developed countries had the lowest infection rates, ranging from 2.9% to 26.8%, with an average of 10.4% ( $SD=0.06$ , 95%  $CI= 0.05-0.07$ ). They were followed by newly industrialized countries, which had infection rates between 6.6% and 54.5% with an average of 26.3% ( $SD=0.10$ , 95%  $CI=0.08-0.12$ ). The highest levels of malware infections were in developing countries, varying from 23.5% to 63.8%, with an average of 38.1% ( $SD=0.10$ , 95%  $CI=0.08-0.14$ ).

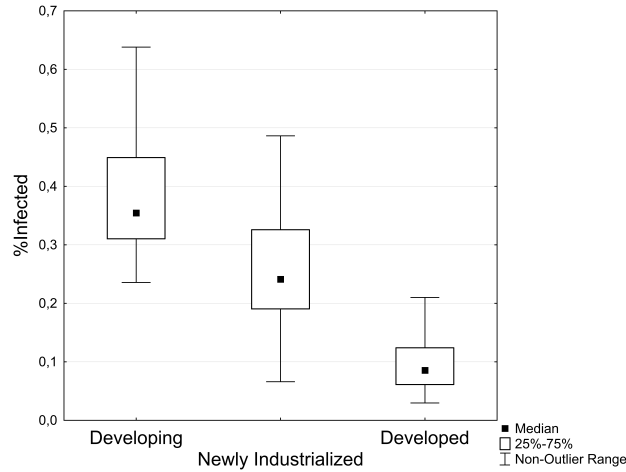


Figure 4.3 Box plot of infection rates by socio-economic status

As for our previous analysis, we first computed the Pearson correlation coefficients to investigate any potential associations between the nine country-level factors and infection rate (see Table 4.6). We further conducted a multiple general linear regression (see Appendix C for detailed results) by stratifying countries based on their socio-economic development to

disentangle the individual effect of each factor.

Table 4.6 Pearson correlation coefficient between infection rate and country-level factors

Factor	Developed			Newly industrialized			Developing		
	r	N	p-value	r	N	p-value	r	N	p-value
GDP	-0.19	35	0.26	0.08	64	0.53	0.35	24	0.10
MYS	-0.68	45	2.10e-07	-0.46	74	4.26e-05	-0.21	26	0.31
FBS	-0.49	43	9.66e-04	-0.17	68	0.16	0.02	24	0.94
%FB(256-2)	0.37	33	0.03	-0.01	30	0.93	0.37	5	0.54
%FB(2-10)	-0.13	36	0.45	-0.06	33	0.76	0.73	4	0.22
%FB(10+)	-0.68	36	6.08e-06	-0.29	24	0.18	-	2	-
IIB	-0.21	43	0.18	-0.40	70	4.91e-04	0.02	25	0.93
%P	-0.81	45	1.29e-11	-0.65	74	3.08e-10	-0.54	26	4.67e-03
GCI	-0.37	44	0.01	0.13	72	0.26	-0.23	26	0.25

### Developed countries

The results from the Pearson correlation in Table 4.6 show that mean years of schooling, fixed broadband speed, fixed broadband subscriptions (above 10Mbit/s), %Protected and the Global cybersecurity index are significantly negatively associated with the infection rate. To the opposite, only the percentage of fixed broadband subscriptions (256kbit/s-2Mbit/s) has a significant positive association with the infection rate.

Five variables (%P, GDP, FBS, IIB, GCI) were identified to be potential risk and protective factors by the regression model for developed countries (see Appendix C). Factors related to cybersecurity (%P, GCI) had, as expected, a negative relationship with the infection rates, which is similar to the results of the global analysis. The quality of technology, in terms of bandwidth and speed, was also found to be negatively correlated with the dependent variable. To the opposite, economic performance (GDP) had a negative sign in the regression model. Surprisingly, mean years of schooling was not found to be a significant factor for developed countries. The insignificance of mean years of schooling can be explained by the higher education in developed countries and less variation of this variable.

Similar to the global model, users' investment in security has the stronger impact on the dependent variable (see Figure 4.2(b)). The second factor appears to be GDP, followed by technology quality (FBS, IIB), and the GCI. Overall, the regression model for developed countries offers a strong predictive ability as it can explain 89% of the variance of the infection rate with the nine country-level factors.

### Newly industrialized countries

The results of the Pearson correlation (see Table 4.6) for newly industrialized countries show that only mean years of schooling, international Internet bandwidth and %Protected are significantly correlated with the infection rate. All factors present a negative association, meaning they could be potential protective factors.

From the regression (see Appendix C), six variables (IIB, MYS, GDP, %P, %FB(2-10), %FB(10+)) were identified to be statistically significant. The results for cybersecurity and education are consistent with our previous findings; they both have a negative association with the dependent variable. While bandwidth presents a negative correlation, broadband speed shows a positive correlation, as opposed to what we previously found. Finally, economic performance (GDP) was negatively associated with the rate of malware infections.

As shown in the Pareto chart (Figure 4.2(c)), the most important factor for newly industrialized countries seems to be bandwidth. Education (MYS) and economic performance (GDP) follow with a similar impact. Users' investment in security (%P) and broadband speed (%FB(2-10), %FB(10+)) are the factors with the smallest effect. Overall, the regression model for newly industrialized countries was able to explain 79% of the variance of the infection rate with the nine country-level factors.

### Developing countries

Based on the results from the Pearson correlation coefficients (see Table 4.6), only %Protected was found to have a significant correlation with the infection rate. Countries with a higher protection coverage (%P) were associated with fewer malware infection rates.

Before conducting the regression we excluded the factors related to fixed broadband subscriptions as data were missing for many developing countries. The results of the regression in Appendix C show that two variables, %Protected and the Global cybersecurity index, were found to be significant in the model. Those variables are both related to cybersecurity and have a negative sign, which means they could be potential protective factors of malware infections. Factors related to education (MYS), economic performance (GDP), and technology quality (IIB, FBS) were not found to be significant for developing countries. This could be explained by lower level of education, economy, and technology for developing countries, resulting in less variation in these variables.

The relative importance of the factors can be visualized by the Pareto chart presented in Figure 4.2(d); users' investment in security has the stronger impact, followed by the GCI. In the end, the regression model for developing countries can explain 41% of the variance of the

infection rate with all six factors. This means that other factors, beyond education (MYS), economic performance (GDP), and technology quality, explain the rate of malware infections for developing countries.

## 4.5 Interpretation

Overall, we found that factors related to economic performance, education, cybersecurity, and Internet connection quality are correlated with the prevalence of malware infection in unprotected hosts. As we discuss below, however, these variables interact in some surprising ways. We also compare our results to those reported in prior studies where possible, and highlight instances in which our findings corroborate or refute theirs.

**Economic performance** While there is some correlation between economic activity (as measured by GDP) and lowered unprotected host infection rates (as measured by Pearson correlation), it appears this relationship is not significant in the global model after controlling for the other variables. Similarly, Garg *et al.* (2013) found no significant association between economic activity (GDP-PPP) and the percentage of total number of spambots, while controlling for other macroeconomic factors. This seems plausible as here factors such as education, technology quality, and cybersecurity investment should explain most of the variance, whereas GDP *per se* should only play a minor role.

When we stratified countries based on their socio-economic status, GDP appeared to be a risk factor for developed countries. This is consistent with the hypothesis that increased GDP means increased incentives for cybercrime, as there are more individuals and organizations with significant wealth to steal from. However, the relationship appeared to be negative for newly industrialized countries; higher economic activity was correlated with reduced malware infection rates. This change of direction may be explained by a potential non-linear association; malware infections decrease as economy grows until a turning point where it rises with economic performance, independently of other risk factors.

A first hypothesis for this relationship could be that GDP acts as a marker for technology quantity, as we removed this factor from our analysis—it was highly correlated with other variables. This would imply that increased ICT adoption in newly industrialized countries could be associated with reduced malware infections, with those countries being more technologically developed and more resilient to malicious attacks than developing nations. However, the effect would be the opposite for developed countries: higher ICT adoption would contribute to increase malware infections, as there are more potential machines to steal from or to exploit for malicious activities (e.g. remote attacks, spam distribution). This explanation

can be tested by examining the partial correlation between GDP and malware infections. In contrast to bivariate correlation, partial correlation allows one to measure the association between two variables while controlling for the effect of other factors. We first computed the correlations while controlling for education, technology quality, and cybersecurity. As expected, the association was positive for developed countries ( $r=0.54$ ,  $N=25$ ,  $p\text{-value}=2.1e-02$ ) and negative for newly industrialized countries ( $r=-0.67$ ,  $N=22$ ,  $p\text{-value}=6.0e-03$ ). We then added the percentage of households with Internet (%HouseholdInternet) to account for ICT penetration. Results show that the associations are still statistically significant for developed countries ( $r=0.55$ ,  $N=25$ ,  $p\text{-value}=2.2e-02$ ) and newly industrialized countries ( $r=-0.67$ ,  $N=22$ ,  $p\text{-value}=9.0e-03$ ), suggesting that ICT penetration cannot explain the non-linear relationship between GDP and malware infections.

A second possibility is that software piracy acts as an intermediate factor between GDP and malware infections. This hypothesis is plausible as software piracy has often been associated with increased risk of malware infections (Mezzour *et al.*, 2015; International Data Corporation, 2013). As economic activity increases, adoption of legal software should also rise (Goel et Nelson, 2009; Andrés, 2006). In contrast, higher economic activity for developed countries could be associated with higher software piracy. While this may be counter-intuitive, Fischer et Rodriguez Andrés (2005) found evidence that software piracy is positively correlated with income for West European and North American countries. To investigate this relationship, we computed the partial correlations between GDP and malware infections while controlling for education, technology quality, cybersecurity, and software piracy. This last factor was collected from the Business Software Alliance and represents the national ratio of the number of unlicensed software units installed to the total number of software units installed for 2011 (Business Software Alliance, 2012). Results show that the associations still hold for developed countries ( $r=0.63$ ,  $N=25$ ,  $p\text{-value}=7.0e-03$ ) and newly industrialized countries ( $r=-0.71$ ,  $N=18$ ,  $p\text{-value}=2.1e-02$ ). This suggests that software piracy may not account for the relationship between GDP and malware infections.

A third explanation could be the distribution of the different versions of Windows (e.g. XP, Vista, 7, 8 and 8.1), i.e., how the operating system (OS) market of a country is shaped could influence his rate of malware infections. To examine this possibility we looked at the distribution of the OS market for unprotected hosts between developed and newly industrialized countries (see Table D.1 in Appendix D). Rates were similar for XP, Vista, and 8.1, but different for 7 and 8. We therefore decided to include the prevalence of Windows 8 (%Windows8), as it is the platform with the highest difference between developed (Mean=10%, SD=4%) and newly industrialized countries (Mean=17%, SD=7%). Partial correlations were computed with education, technology quality, cybersecurity, and prevalence

of Windows 8 for unprotected hosts as control variables. This time, both the associations for developed ( $r=0.26$ ,  $N=25$ ,  $p\text{-value}=2.6e-01$ ) and newly industrialized countries ( $r=-0.36$ ,  $N=22$ ,  $p\text{-value}=2.0e-01$ ) were found to be not statistically significant. This suggests that the OS market distribution may be an intermediate factor between GDP and malware infections. Although our analysis provides empirical support for this explanation, it is necessary to develop and test new theories that can account for the causes of this relationship.

Overall, results suggest that GDP *per se* is not a significant factor of malware infections. Rather, economic performance would act as a distal factor via multiple intermediate variables (e.g. technology quality, OS market distribution) that were captured in our analysis.

**Education** Education seems to be more consistently associated with reduced malware infection rates. As expected, mean years of schooling was negatively correlated with malware infections in the global analysis. When we stratified countries by socio-economic status, education was only significant for newly industrialized countries. Similarly, Microsoft Burt *et al.* (2014) found that countries with high education, as measured by the literacy rate, are less likely to be infected by malware.

Overall, our analysis suggests that education is a significant distal factor of malware infections. This could imply that education is involved in the causal chain via a number of intermediate factors (e.g. IT literacy, cybersecurity awareness) that were not captured by our analysis. Another potential explanation is that mass education, as measured by mean years of schooling, has a direct aggregate effect at the population level. Testing those hypotheses would require the collection of more specific data on potential intermediate factors, both at the population and individual level. From there, additional studies could be designed to disentangle the aggregate effect of education, if any, from its indirect effect on malware infections.

**Technology** The quality of a country's technological infrastructure does seem to be correlated with reduced malware infections. Increased international Internet bandwidth and high fixed broadband speed were both associated with reduced unprotected host infection rates when looking at the bivariate correlations. After controlling for economic development, education level, and cybersecurity investment, bandwidth was found to be a protective factor, regardless of the socio-economic development level. The effect of broadband speed in terms of direction and magnitude, however, turned out to depend of the socio-economic status. While higher broadband speed (FBS) was negatively correlated with malware infections for developed countries, higher proportions of moderate (%FB(2-10)) and high speed

fixed broadband (%FB(10+)) were actually positively correlated with infections for newly industrialized countries.

One explanation for this inconsistent relationship between the quality of Internet connectivity and infection rates is that while better bandwidth makes it easier to keep systems updated, faster connections make it easier for attackers to exploit large populations of unmaintained systems. To partially investigate this hypothesis, we first looked for associations between measures of system status and bandwidth. We computed from MSRT the percentage of users that had out of date AV signatures during the 4-month period and the percentage of users who performed their Windows updates every month during the study. As expected, the first measure (%Out-of-date AVs) was negatively correlated with bandwidth ( $r=-0.59$ ,  $N=186$ ,  $p\text{-value}=4.47\text{e-}15$ ) while the second measure (%Up-to-date Systems) was positively correlated ( $r=0.75$ ,  $N=186$ ,  $p\text{-value}=3.13\text{e-}28$ ). This suggests that bandwidth could be a protective factor for malware infections though various measures of system status as intermediate variables. However, testing the second part of the hypothesis —that faster connection makes it easier to infect large populations of vulnerable computers— would require conducting large-scale studies of malware propagation based on epidemiological models.

Overall, these findings provide evidence that bandwidth could be a protective factor that contributes to decreased risk of malware infections via multiple intermediate variables related to system status. Moreover, results suggest that fast broadband connections are associated with reduced malware infections in only some circumstances. Further studies are required to determine the exact nature of the causal relationship, if any.

**Cybersecurity** Individual investment in security (as measured by %Protected) appeared to have a strong negative correlation with malware infections for all countries, regardless of their socio-economic status. Intuitively, countries with higher percentage of users protected by antivirus products were found to have lower unprotected host infection rates. A first explanation for this observation is that antivirus product penetration acts as a marker for other variables that were not captured by our analysis. For example, usage of antivirus products may be related to individual risk-taking behavior —users who tend to underestimate cybersecurity-related risk may tend to unprotect their computer. Hence, AV penetration could be a marker for risk-attitude towards cybersecurity at the population level. One potential way to investigate this hypothesis would be to correlate AV market penetration with individual risk-taking behavior in other specific contexts, such as finances, sports and leisure, health, career, and car driving Dohmen *et al.* (2011). As a first attempt, we correlated %Protected with tobacco consumption. We used 2012 male smoking prevalence among persons aged 15 years and over from the World Health Organization (World Health Organization,



2015) as an aggregate measure of risk attitude in the domain of health (Hersch et Viscusi, 1990; Dohmen *et al.*, 2011; Feinberg, 1977). We first performed a bivariate correlation using Pearson correlation coefficient to investigate any linear association between the two variables: results indicate that the association is not significant ( $r=-0.019$ ,  $N=104$ ,  $p\text{-value}=0.845$ ). As smoking is a function of various determinants (e.g. education, income, social support) beyond risk attitude, we also performed a partial correlation. We used the mean years of schooling (MYS) and GDP as markers of socio-economic status. This time, results of the partial correlation reveal a weak negative association between %Protected and smoking prevalence ( $r=-0.22$ ,  $N=92$ ,  $p\text{-value}=0.036$ ), suggesting that AV penetration may be a marker of risk-taking behavior towards cybersecurity at the population level. Although our analysis provides limited empirical support, validation of this hypothesis would require to conduct either country level studies based on aggregated measures of cybersecurity risk-aversion or large-scale user studies. A second but tenuous possibility is that unprotected systems benefit indirectly from the protection of other —protected— systems. This is similar to the free-rider effect in economics, where non-paying individuals can benefit from the goods, resources or services of others, even though they did not pay for them. Unprotected hosts would then benefit from a “AV herd immunity” effect from systems protected by antivirus products. Even though prior work (Blackbird et Pfeifer, 2013a; Lalonde Lévesque *et al.*, 2015) have provided some empirical evidence for this explanation, proper validation should be achieved by conducting further epidemiological studies designed for the purpose.

The Global cybersecurity index was found to be a weak protective factor for developed and developing countries —its effect was not significant in the global model and for newly industrialized countries. In comparison, previous work (Kleiner *et al.*, 2013; Png *et al.*, 2008; Van Eeten *et al.*, 2010) provided limited empirical evidence of the effect of national policies on cybersecurity. However, those results can’t be directly compared to our research, as previous studies used various proxy variables to evaluate the impact of cybersecurity policies. Overall, our results tend to confirm that investment in cybersecurity at the national level, as measured by the GCI, is associated with reduced unprotected host infections. From there, further studies could be conducted to understand the individual contribution of each component of the GCI (legal measures, technical measures, organizational measures, capacity building or cooperation) and help design better evidence-based cybersecurity policies.

## 4.6 Study limitations

This study and its conclusions are subject to a number of limitations and potential bias. First, there is an inherent limitation to our results because our sample population is drawn from

Windows systems running MSRT; thus, it does not provide insight into Windows systems that do not run Windows Update, and it does not give insight into the infection rates on non-Windows systems such as MacOS and Unix-based OSes. Furthermore, the analysis was limited to personal computers (e.g. desktop and laptop) meaning that the factors identified may differ significantly on mobile devices and tablets. However, given that there are more than one billion computers regularly running MSRT, patterns discovered in this population are important on their own, whether or not they are representative of patterns in other computational contexts.

Another significant limitation is that the detections from MSRT are only for a subset of malware families. While these families may represent some of the most significant malware families on Windows, they are not representative of the entire threat landscape, and so MSRT reported infection rates will be different from the overall malware infection rate. Nevertheless, given the significance of MSRT-targeted malware these infection rates are also of inherent interest. Detections by MSRT are also dynamic and fluctuate over time (Burt *et al.*, 2014). To partially compensate this volatility, we used the *period prevalence* of malware infections, that is the prevalence during a specific period of time. While period prevalence may be a better measure than averaged prevalence, our measurement may still be subject to temporal variation, as is often the case for security data (Edwards *et al.*, 2015). Moreover, malware infection rates reported in this study may not be representative of other time frames.

As this was an observational study at the population level, we only intended to identify correlations to generate hypotheses on the causes of malware infections. We did not attempt to infer causation. While ecological studies can be used to identify potential factors based on aggregate variables, care must be taken to avoid the risk of ecological fallacy—an error in the interpretation of the results when conclusions are inappropriately made about individuals based on aggregated data. The fallacy assumes that individual members of a group all have the average characteristics of the group. Another limitation of ecological studies is their susceptibility to confounding. Both economic and education factors have been considered control variables in our study. We cannot ensure, however, that our results are not affected by other unknown extraneous variables. Although we included a broad set of national-level factors in our study, there may be other plausible predictors of malware infections that were not captured through our analysis. It would be interesting in future work to consider additional factors, such as culture, demographics, technology quality, or private investment in security, as the latter two were excluded due to high multicollinearity.

## 4.7 Conclusion and policy implications

We presented the results of the first ecological study applied to computer security designed to identify national-level malware infection risk factors. We found relationships with economic performance, education, Internet connectivity, and cybersecurity that have not been previously reported, particularly in how their relationships with infection rates are not simple correlations. We also explored in detail the potential underlying causes between the studied national-level factors and malware infections.

While our work corroborates some findings in earlier research, our results suggest that GDP *per se* is *not* a significant factor of malware infections. Rather, economic performance could be a distal factor acting through multiple intermediate variables, such as technology quality, OS market distribution, or education. The later, as measured by mean years of schooling, was also found to be a protective distal factor of malware infections. However, the question of whether it is a direct aggregate effect or an indirect effect should be investigated in further studies. We also found evidence that bandwidth acts as a protective factor of malware infections via multiple variables of system status. Interestingly, results suggest that Internet connection quality, as measured by broadband speed, may be a protective factor in only some circumstances. While high broadband speed was associated with reduced malware infections for developed countries, its effect was the opposite for newly industrialized countries. The percentage of AV-protected machines and the Global cybersecurity index were also found to be significant protective factors. This suggest that investment in cybersecurity, both at the individual and national level, could contribute to reduce the risk of malware infections. Finally, our work shows that risk and protective factors may not have the same effect and relative importance in countries with different socio-economic status.

More interestingly, our findings have potential policy implications. For example, education was identified as a major protective factor —countries with higher level of education had lower malware infection rates. Although education was measured at the population level, this may suggest that governments should prioritize investments in user education. Such efforts could focus, among others, on the promotion of safe computer behavior, like installing an AV product, or keeping applications, software and OS up-to-date. However, although user education may foster the adoption of safe computer behavior, it is possible that many risky computer behaviors, particularly in developing and newly industrialized countries, are also determined by income. For example, users in such countries may face a trade-off between buying a legitimate software or downloading a pirated software and saving money. Prior understanding of how risky computer behavior is determined by a lack of cybersecurity risk awareness and the costs of adopting safety measures and behaviors would therefore be

essential in the success (or failure) of such interventions.

Similarly, technology quality (as measured by bandwidth) was also identified as a protective factor. While users can install free AV products, they will not be fully effective if their signature databases cannot be updated as a result of poor Internet connection. This could also suggest that governments should invest in ICT infrastructure. On the other hand, investing in better ICT infrastructure on the basis of risk reduction alone might not represent a sufficiently great value proposition for developing and newly industrialized countries. Moreover, we found evidence that higher broadband speed was associated with higher malware infection rates for newly industrialized countries, while its effect was the opposite for developed countries. Hence, interventions proven to be successful in developed countries might not be effective (or even possible due to budget constraints) in newly industrialized and developing countries.

In light of this discussion, we believe that policy interventions, whether technical, legal, or educational, might prove ineffective if they do not take into account the socio-economic circumstances of populations and individuals. As shown by our findings, it is important to consider the socio-economic status of countries in future risk analysis of security threats and consequent evidence-based decision making. Moreover, the relative effect of protective factors were found to differ depending on the socio-economic context. This suggests a prioritisation of efforts by policy makers, where stronger protective factors should be leveraged first. For example, for newly industrialized countries it would seem that increasing bandwidth availability would have stronger effect than increasing AV usage. In contrast, the opposite is true for developed countries. In both cases, however, this prioritisation of effort must also take into account the relative cost-effectiveness of such counter-measures, e.g. can a similar effect be more effectively obtained by investing the same amount of resources to address one factor vs. the other. Assessing the cost-effectiveness of such counter-measures is beyond the scope of our study.

This work also demonstrates that rigorous ecological studies can be used to identify risk factors for malware infections at the population level. We believe a population health approach could provide a skeleton from which security threats can be researched and for which appropriate national-level interventions can be developed. It is important that further research be conducted to assess the multi-level risk factors of malware infections, in order to verify some of the hypotheses we have advanced and establish sound causation. Since explaining individual cases requires that we consider both underlying causes of infection in the population and individual circumstances, research into the different levels of risks should be seen as complementary. From there, cybersecurity policies could be designed to reduce the prevalence of malware considering both individual and ecological influences. We hope this work illustrates

the merits of future large-scale ecological studies applied to computer security.

**Acknowledgements** The authors would like to thank the Microsoft Malware Protection Center (MMPC) for granting us access to the MSRT telemetry data and for supporting this work. First author would also like to thank Thierry Lavoie for his useful comments and suggestions on designing the study in the paper. Finally, we would like to thank our anonymous reviewers who provided many helpful comments on the paper.

## CHAPITRE 5    ARTICLE 2 : TECHNOLOGICAL AND HUMAN FACTORS OF MALWARE ATTACKS : A COMPUTER SECURITY CLINICAL TRIAL APPROACH

*Published in June 2018 in ACM Transactions on Security and Privacy.*

**Authors**    Fanny Lalonde Lévesque<sup>1</sup>, Sonia Chiasson<sup>2</sup>, Anil Somayaji<sup>2</sup>, José M. Fernandez<sup>1</sup>

**Institutions**    École Polytechnique de Montréal<sup>1</sup>, Carleton University<sup>2</sup>

**Abstract**    The success (or failure) of malware attacks depends upon both technological and human factors. The most security-conscious users are susceptible to unknown vulnerabilities, and even the best security mechanisms can be circumvented as a result of user actions. Although there has been significant research on the technical aspects of malware attacks and defence, there has been much less research on how users interact with both malware and current malware defences.

This paper describes a field study designed to examine the interactions between users, antivirus (AV) software, and malware as they occur on deployed systems. In a fashion similar to medical studies that evaluate the efficacy of a particular treatment, our experiment aimed to assess the performance of AV software and the human risk factors of malware attacks. The 4-month study involved 50 home users who agreed to use laptops that were instrumented to monitor for possible malware attacks and gather data on user behaviour. This study provided some very interesting, non-intuitive insights into the efficacy of AV software and human risk factors. AV performance was found to be lower under real-life conditions compared to tests conducted in controlled conditions. Moreover, computer expertise, volume of network usage, and peer-to-peer activity were found to be significant correlates of malware attacks. We assert that this work shows the viability and the merits of evaluating security products, techniques and strategies to protect systems through long-term field studies with greater ecological validity than can be achieved through other means.

**Keywords**    Antivirus, Malware, Field study, User study, User behaviour, Security behaviour, Clinical trial, Risk factor, Human factor, Security product, Usability

## 5.1 Introduction

Malicious activity on the Internet is continuously evolving; the nature of threats changes rapidly. Modern malware authors adapt their techniques to exploit new vulnerabilities, take advantage of new technologies, and evade security products. Users may be enticed to take direct (or indirect) actions that lead to the infection of their computers. Some actions, such as opening an email attachment or visiting a malicious web site may occur immediately prior to infection. Others, such as not updating system or willingly installing software whose true intention is masked, may occur over time so that a combination of actions lead to a vulnerable system state.

Meanwhile, antivirus (AV) products have evolved in response. The signature-based file-scanning engines that used to be the core technology of AV products have been complemented by multiple layers of protection, including identification of hazardous URLs, reputation-based software classification and system behaviour monitoring (Saeed *et al.*, 2013). Computers are no longer stand-alone machines that need to be protected as such, and what used to be a security problem—their connectedness—is increasingly being leveraged by AV vendors to better protect their customers. Periodic signature file updates are being replaced by on-demand resource lookups on databases in cloud infrastructures; these databases are in turn fed by the continuous reporting of millions of AV client installations (Saeed *et al.*, 2013; Alam *et al.*, 2014). AV products have thus evolved into complex “anti-malware” software, or rather complex software systems involving several semi-independent components with which the user must occasionally interact. While many AV vendors try to make the installation and operation of their product as usable as possible, the truth is that the AV’s operation and performance still depends on the user. This dependence is due to user configuration of the many AV features and to other user-driven factors such as how often the machine is connected to the Internet, how often its software and signatures are updated and, most importantly, how users are interacting with the computer and the Internet when confronted with situations where their actions could lead to infection.

In other words, the operating environment of both AV products and malware not only includes the machine they are trying to protect/penetrate, but also the network that connects it to the rest of the world and its *user*. Indeed, the human is part of the operating environment of the machine, along with the software that attempts to execute on it or protect it. It thus seems natural to adopt a *Human in the loop* approach to evaluate the performance of AV products and the susceptibility of users to getting their machines infected. This change of paradigm is fundamental if we want to better understand what role users really play in the process of malware infection. In particular, it becomes paramount to understand how human

factors, such as demographics, computer literacy, perception of threat, and user behaviour may affect the risk of malware infection.

This philosophy of *Homo in machina* is also in sharp contrast with current AV evaluation methods, which are largely based on automated tests performed in controlled environments (Gordon et Ford, 1996; Marx, 2000; Harley, 2009; Edwards, 2013). While these tests are adequate to evaluate AV products under specific scenarios, they do not measure the “real world” efficacy of AV products as deployed on machines operated by real users (Lalonde Lévesque *et al.*, 2016a). Even the most advanced tests, which include automated user profiles (PC Security Labs, 2013), cannot accurately capture all user behaviour and other external factors, such as evolving malware threats or different system configurations, that may affect how AV perform.

To address these shortcomings and to understand the influence of human factors in malware attacks, an alternative method, *computer security clinical trials*, was proposed in 2009 by Somayaji *et al.* (2009). We conducted the first such experiment at the École Polytechnique de Montréal in 2011-2012, involving 50 home users using their own computers in everyday life for 4 months. This journal version of our work details the methodology (Lalonde Lévesque *et al.*, 2012a; Lalonde Lévesque et Fernandez, 2014) and extends the preliminary results already presented in earlier work both in terms of AV product evaluation (Lalonde Lévesque *et al.*, 2012b) and human risk factors (Lalonde Lévesque *et al.*, 2013). In particular, this paper i) provides a new evaluation of the AV product including the efficacy (Section 5.5.3) and the user experience (Section 5.5.4), ii) reinterprets and updates statistical analysis on human factors (Section 5.6.1), iii) investigates some unstudied user behaviour factors (Section 5.6.2), such as system activity, applications usage, network usage, files downloaded, peer-to-peer activity, and level of vigilance, and iv) extends the discussion to include the implications of these new results.

The remainder of the paper is structured as follows. In Section 5.2, we present related work in AV product evaluation, and human factors related to computer threats. Section 5.3 details the concept of computer security clinical trials. Section 5.4 describes its methodology. In Section 5.5, we discuss the results of the study in terms of threat detections by AV, missed detections, and user experience. Section 5.6 identifies potential risk factors related to user characteristics, demographics and behaviour. We discuss limitations of our work in Section 5.7. Finally, we conclude and summarize the results and implications of this work in Section 5.8.



## 5.2 Related work

Numerous studies evaluating the performance of AV products and the influence of human factors on information technology (IT) security have been conducted in recent years. We describe in Section 5.2.1 the current state of affairs regarding AV product evaluation and we present in Section 5.2.2 previous research related to humans factors (user characteristics, demographics and behaviour) and computer threats.

### 5.2.1 Antivirus product evaluation

Antivirus products offer important information system protection against current threats. Testing how these products are effective at protecting end-users and their systems is therefore crucial. We present here a critical review of current evaluation methods and discuss their limitations.

**Controlled conditions** Typical evaluations by commercial testing labs (e.g. AV Comparatives (2013); PC Security Labs (2013)) are based on automated tests conducted in controlled environments. For example, file scanning tests (also called “static” or “on-demand”) are based on scanning collected or synthesised malware samples along with legitimate programs. As there is no file execution, i.e. there is no software behaviour to analyze, static tests cannot adequately reflect the performance of products using active and proactive detection. Dynamic tests consist either of executing files or exposing antivirus products to known bad URLs (Edwards, 2013). While this latter type of tests does evaluate the performance of AV products as a whole (and not that of individual features), they may not be representative of what is typically experienced by users “in the wild”.

One major issue with in-lab tests is that the sample malware collection is often too small, inappropriate, and not validated (Harley et Lee, 2008; Kosinar *et al.*, 2010); this problem is often referred to as the *sample selection problem*. This can easily bias the test results, whether consciously or not, and can thus severely limit their usefulness. To this end, the WildList Organization International has proposed in 1993 the WildList (The WildList Organization International, 2017), a cooperative listing of malware. This list, which only contains malware observed in the wild, has the main advantage of being validated by security professionals. However, it may not be representative of the most common malware in real-time, as it is only updated monthly. To partially address this shortcoming, the Anti-Malware Testing Standards Organization (AMTSO) created in 2013 the Real Time Threat List (RTTL) to provide a real-time view of threats as they are found in the wild (Zwienenberg *et al.*, 2013).

The RTTL allows testers to conduct evaluations based on malware samples that represent the current state of the malware ecosystem. Although AV tests against such data sets are more realistic, they are not ecologically valid in that the effect of the human factor is not being measured.

Some researchers (Muttik et Vignoles, 2008; Vrabec et Harley, 2010) suggested emulating user interaction with scripts and creating user-specific testing scenarios. For example, testing for a gaming profile should prioritize network latency or reduction in frame rate, while testing for a worker profile should emphasize on downloading files from a server or audio/video file editing. As a first attempt, PC Security Labs conducted in 2013 an AV test (PC Security Labs, 2013) to measure the defence efficiency of AV solutions against seven different types of user profiles: Internet addict, network businessman, socializer, basic user, gamer, self-presenter and infrequent user. Their test confirmed that AV solutions perform differently depending on the user profile. However, though this testing approach simulates a more realistic operational conditions, it is impossible to capture all user behaviour, and external factors that may affect AV efficacy in real-life.

**Real-life conditions** One complementary approach to tests performed under controlled environments is to conduct evaluations in an holistic environment where the system, the AV product, and the user are included. For example, some observational field studies of AV products have been conducted. In such tests, AV products are not installed on systems. Rather, systems are monitored with their actual protection without any intervention. Blackbird et Pfeifer (2013b) used data from the Malicious Software Removal Tool (MSRT) on millions of systems to evaluate how AV protection state impacts infection rates. Lalonde Lévesque *et al.* (2015) also used MSRT data to measure the overall performance of the AV ecosystem over a four-month period. In another study from Lalonde Lévesque *et al.* (2016a), the authors used data collected from the MSRT and Microsoft’s Windows Defender on millions of systems to conduct a large-scale comparative test of AV products. Their findings showed that AV performance varies significantly as a function of external factors, such as user factors, environmental factors, and malware types.

Another potential way to assess AV performance in real-life is to conduct experimental field studies. For example, Somayaji *et al.* (2009) proposed in 2009 conducting computer security clinical trials inspired by the same methodology as used in medical trials. In this method, security products are randomly deployed on specific populations and are monitored to assess their real world performance in normal use. However, to the best of our knowledge, there has been no such studies of AV products published in the literature other than our previously published work (Lalonde Lévesque *et al.*, 2012b; Lalonde Lévesque *et al.*, 2014).

### 5.2.2 Human factors and computer threats

In this section, we present a review of past work that studied how human factors, such as user demographics, characteristics and behaviour, correlate with computer threats.

**Subjective research methods** One approach to studying the human factors in computer security is to adopt a subjective research method. This type of approach seeks to explore the perception and the attitude of users when they are facing computer security decisions. It primarily uses qualitative methods such as surveys, interviews and observations to understand how and why participants interact with computer systems.

For example, Milne *et al.* (2009) applied protection motivation theory and social cognitive theory to understand online customers' risky behaviour and protection practices. They conducted a national online survey of 449 non-student respondents in 2009 and confirmed that age and gender are significant correlates of online risky behaviours; males and younger users were found to be more likely to adopt risky behaviours online. This assertion was also confirmed in 2010 by Sheng *et al.* (2010) who conducted an online study with 1001 users to evaluate their susceptibility against phishing attacks. They concluded that prior exposure to phishing education is associated with less susceptibility to phishing, suggesting that phishing education may be an effective tool. They also found that age is a contributing risk factor and that young people aged between 18 to 25 are more susceptible to phishing. Using a sample of 295 college students, Ngo et Paternoster (2011) applied the general theory of crime and lifestyle/routine activities framework to assess the effects of individual and situational factors on seven types of cybercrime victimization, including computer virus infection. They conducted a self-assessment survey in 2011 and deduced that age is a significant risk factor for computer virus infection, with older respondents being less likely to get infected. In another study, Bossler et Holt (2009) applied a routine activities framework to explore the causes and correlates of self-reported data loss from malware infection. The authors administered a survey on a sample of 788 college students in 2006 and investigated, among others, the effect of gender, age, race, and employment status. They found that being a female and being employed increases the odds of data loss compared to male and unemployed users respectively. However, age was not identified as a significant predictor of self-reported data loss from malware infection. Similarly, Reyns *et al.* Reyns (2013) also applied the routine activity theory to study online crime. Using data from a sample of 5,985 participants from 2008 to 2009, they investigated the relationship between individual's online routines, characteristics (age, gender, employment, income) and identity theft victimization. Results suggested that age, gender, employment, and income were significant correlates, where older respondents,

males, employed respondents, and those with higher incomes were more at-risk. The authors also found that using the Internet for banking, shopping, communicating (e-mail, instant messaging), and downloading, is associated with increases in the likelihood of identify theft. Onarlioglu *et al.* (2012) conducted a survey in 2011 on 164 Internet users who possess diverse backgrounds and varying degrees of computer security knowledge. Results confirmed the general intuition that technical security knowledge has a considerable positive impact on user ability to assess risk, especially when the threats involve technically complex attacks. Finally, Grimes *et al.* (2007) surveyed 207 participants in 2007 to study how computer-related characteristics, online behaviours, and demographics (age, gender) correlate with spam attitudes and actions. The authors found no significant association between demographics and self-reported reception of spam. However, they did find some evidence linking specific online behaviours, such as purchasing online, making a web page, or posting in a newsgroup, and self-reported reception of spam.

**Objective research methods** Another complementary approach to study human factors is to conduct studies based on an objective research method. While subjective studies will allow researchers to better understand user perception and perspective regarding computer threats, an objective method, either based on qualitative or quantitative data, will allow to study and measure user behaviour regarding computer security. For example, one approach to identify potential risk factors related to malware infection is to conduct observational or experimental studies based on real-life data, as self-reported data may lack ecological validity to represent actual user behaviour.

Maier *et al.* (2011) performed in 2011 an empirical study based on network traces from residential users to analyse the relationship between security hygiene (AV and OS software updates) and potential risky behaviour. They found that computer hygiene has little correlation with observed behaviour, but that risky behaviour, such as accessing blacklisted URLs, can more than double the likelihood that a system will manifest security issues at the network level, e.g. sending spam, performing address scans or communications with bot-net command-and-control (C&C) servers. Canali *et al.* (2014) performed a comprehensive study on the effectiveness of risk prediction based on the web browsing behaviour of users in 2013. Their results showed that the more websites a user visits, the higher is his exposure to threats. Ovelgönne *et al.* (2017) leveraged 2009-2011 telemetry data from the Symantec's Worldwide Intelligence Network Environment (WINE) project (Dumitras, 2017) to study the relationship between user behaviour and cyber attacks. They created 4 user profiles (gamers, professionals, software developers, and others), and studied how 7 machine features (number of binaries; fraction of unsigned, downloaded, low prevalence, and unique binaries; number of

ISPs to which the user connected) correlate with the number of attempted malware attacks by host machine. The authors found all features to be significant contributing factors, suggesting that heavy downloading of binaries, traveling a lot, and downloading rare pieces of code could increase the risk of malware attacks. In addition, they found software developers to be the most prone to malware attacks.

Some researchers have focused on phishing susceptibility. Jagatic *et al.* (2007) launched in 2005 a real (harmless) phishing attack targeting 581 university students to quantify how reliable social context would increase the success of victimisation. Through their analysis, they found that females were more likely to fall victim of the social phishing attack. The attack was also slightly more successful with younger targets. Kumaraguru *et al.* (2009) conducted in 2008 a real-world study to evaluate phishing training effectiveness, and investigate how users' demographic factors influence training and phishing susceptibility. Their results showed no significant difference between males and females. However, they found participants in the 18-25 age group to be consistently more vulnerable to phishing attacks than older participants. In another study, Oliveira *et al.* (2017) investigated spear phishing susceptibility as a function of user age, gender, weapon of influence (scarcity, authority, commitment, etc.), and life domain (financial, health, social, etc.). The authors performed a 21-day study involving 158 participants, which took place in the participants' homes from 2015 to 2016. After exposing participants to experimentally controlled spear phishing emails, researchers found that women, particularly older women, were more susceptible to phishing attacks. Moreover, their results highlighted the extent to which younger and older participants differ in their susceptibility to various weapons of influence (scarcity, authority, commitment, etc.).

Other studies have adopted a methodological approach based on the concepts and methods of epidemiology. This approach refers to the likely causes and risk factors for infection, understanding the spread of malware and, where appropriate, the methods to remedy it. For example, Carlinet *et al.* (2008) designed a case-control study in 2006 to analyse the behaviour of ADSL customers and identify customer characteristics that are risk factors for malware infection. The study showed that using the Windows operating system and heavily using web applications and streaming are major risk factors of malware infection. Lee (2012) also conducted in 2010 a case-control study of academic malware recipients to identify putative factors that are associated with targeted attack recipients. The experiment revealed that specific individual profiles, such as individuals working in Eastern, Asiatic, African, American and Australian Languages, Literature and Related Subjects and Social Studies, especially Economics, are at a statistically significant elevated risk of being subjected to targeted attacks compared with others. Following the same methodology, Thonnard *et al.* (2015) designed a case-control study to identify organizational and individual risk factors of

targeted attacks. Based on a large corpus of targeted attacks blocked by an email scanning service from 2013 to 2014, they showed that directors and high-level executives are more likely to be targeted, and that specific job roles such as personal assistants are even more at risk of targeted attack compared to others. Lalonde Lévesque *et al.* (2017) conducted another case-control study specifically designed to evaluate the independent effect of age and gender on the risk of malware victimisation. Using data collected from Microsoft’s Windows Defender on a sample of three million devices in 2015, the authors found that both age and gender are significant contributing factors for malware encounters. Men, and young men in particular, were found to be more susceptible to malware attack than women, and younger users to be more at risk than their older counterparts. Interestingly, results also suggested that the effect of age and gender is not constant across different types of malware; women were slightly more susceptible to encounter adware, and older users were more susceptible to rogue malware and ransomware. Also inspired by the epidemiology approach, Yen *et al.* (2014) conducted in 2013 a study of malware encounters in a large, multi-national enterprise. They coupled malware encounters with Web activities and demographic information. Their results suggested that user demographic and behaviour features can be used to infer the likelihood of malware encounters; males and people with technical expertise were found to be more likely to encounter malware.

### 5.3 Computer security clinical trials

One potential way to study technological and human factors of malware attacks is through conducting clinical trials of software, as proposed in 2009 by Somayaji *et al.* (2009). With such clinical trials, security software is installed and monitored on systems in regular use by regular users. Data is then gathered on the performance of the security software in protecting the system and on how the user interacted with the system during this time period. By correlating user behaviour, application use, and security software activity, we can gain insights into the interactions between all three in an ecologically valid context.

For this first experiment, we evaluated one single AV product and we fixed some of the external factors that could affect a computer’s likelihood of being infected by malware. For instance, all users were selected in the same geographic area. They all had the same laptop and system configuration, as those factors could affect the AV performance in protecting the system. The main reason behind such decisions was to minimize the number of free variables and reduce the complexity of designing, conducting and analysing the results of this first proof-of-concept study. Moreover, the data collected during the experiment considered many of the other reasonable factors that could influence malware attacks such as user profiling,

user behaviour, host configuration and environment.

## 5.4 Study description

This first experiment of its kind was conducted from November 2011 to February 2012 as a proof-of-concept study involving 50 participants. The study monitored real-world computer usage through diagnostics and logging tools, monthly interviews and questionnaires, and in-depth investigation of any potential infections. The study had the following goals:

1. Develop an effective methodology to evaluate AV products in a real-world environment;
2. Determine how malware infects computer systems and identify source of malware infections;
3. Determine how phenomena such as the system configuration, the environment in which the system is used, and user behaviour affect the probability of infection of a system.

### 5.4.1 Ethics clearance

The project was examined and cleared by the two relevant university entities: the *Comité d'évaluation des risques informatiques* (CERI, i.e., the computer security risks evaluation committee) and the *Comité d'éthique de la recherche* (CER, i.e., research ethics review committee).

### Computer risks

We provided users with an AV product that was centrally managed on our own server to guarantee high-availability. The AV software was updated daily and configured to perform a full scan of the computer every day, to provide an equal or better level of protection than average corporate or home users would normally have. Should the AV detect an infection, it would be automatically neutralized. Conversely, in the event that our diagnostics tools detected an infection on the computer that had been undetected by the AV, a procedure was given to users so they could neutralize the threat by themselves.

Giving that the experiment implied manipulation of malware files, special precautions were taken to protect the university IT infrastructure. All malicious or potentially malicious files were first encrypted and copied to DVDs before being stored in the high security zone of the laboratory. Moreover, all computers were analysed by being connected to an isolated network to prevent any contamination of the university network.

## Ethical and privacy considerations

Following the computer security risks evaluation committee clearance, the research ethics review committee cleared our recruiting procedures, the experimental protocol, as well as the measures adopted for user anonymity and confidentiality of the data collected.

To ensure the anonymity of users, we assigned each user a unique identification (ID) number associated with his computer. The only personal information kept for administrative and financial purposes was the participant's name, email address, and telephone number. This information was only accessible by the project leader and was destroyed 3 months after the end of the study. All raw data and statistics generated during the experiment were sanitized. The data was stored in a locked cabinet in the high-security zone of the laboratory, which is protected with three-factor authentication (biometrics, PIN, and ID card). This work zone is completely isolated from the Internet and the university network. The security policy of the laboratory was also applied to the deletion of all personal data related to the experiment. This policy applies to all information whether on paper or electronic media, and conforms with Government of Canada information security standards.

Only authorized personnel within the context of the project was able to access the data. In the event we wanted to share the anonymized data with other researchers, they had to agree to comply to the university computer risks and ethics policy. Moreover, the data collection was bound to the purpose of the project's research objectives. Finally, if we had inadvertently discovered information leading a reasonable person to believe that a (serious) crime had been committed or was about to be committed, we would have been required by law to advise the appropriate authorities (law enforcement agencies, etc.). Fortunately, this was not the case in this experiment.

### 5.4.2 Equipment

The laptops provided to the subjects all had identical configurations, with the following software installed: Windows 7 Home Premium; Trend Micro's OfficeScan 8.0; monitoring and diagnostic tools including HijackThis, ProcessExplorer, Autoruns, SpyBHORemover, Spy-DLLRemover, tshark, WinPrefetchView, WhatChanged; and custom Perl scripts developed for this experiment. These tools and their use in our experiment are described in Section 5.4.3.

Scripts were used to automate the execution of the tools and compile statistical data about system configuration, the environments in which the system was used, and the manner of use. The data compiled by our scripts included:

- The list of applications installed;



- The list of applications for which updates were available;
- The number of web pages visited per day;
- The number of web pages visited by categories per month;
- The number and type of files downloaded from the Internet;
- The number of different hosts to which the laptop communicated;
- The list of the different locations from which the laptop established connection to the Internet;
- The number of hours per day the laptop was connected to the Internet;
- The number of hours per day the laptop was powered on.

Before deployment, we profiled the laptops to establish a baseline data set in order to compare at later date the variation in infection rates induced by AV and hardware choices vs. that generated by variation in demographics, behaviour and software configuration. The recorded information included: i) a hash of all files plus information about whether the files were signed; ii) a list of auto-start programs; iii) a list of processes; a list of registry keys; a list of browser helper objects (BHO); iv) a list of the files loaded during the booting process; and v) a list of the pre-fetch files.

The AV product was centrally managed on our own server, in a manner similar as is usually done for corporate installations to centralise distribution of signature file updates. All AV clients installed on the laptops were thus sending relevant information to our server about any malware detected or suspected infections as they occurred.

### 5.4.3 Experimental protocol

#### Subject recruiting

We recruited by advertising the experiment on the Université de Montréal campus (which includes the École Polytechnique engineering school and the HEC business school) using posters and newspapers. Even though the recruiting process was centered on the university campus, the study was open to everyone. Interested participants were invited to visit a designated Web site to obtain more details and fill a short on-line questionnaire that we used to collect initial demographic information such as gender, age, status and field of expertise. The only inclusion criteria was to be at least 18 years old.

Given our limitation on study sample size (number of laptops available), an important issue was to select a sample of 50 users who were as representative as possible of the general population of Internet users. Due to the over-representation of students and the limited number of candidates, we selected users based on a cluster sampling technique where users

were grouped by their demographic characteristics. While this approach was suitable for a first study, recruiting for larger-scale trials should be more rigorously structured, as is the case for medical clinical trials.

### **In-person sessions**

Users were required to attend 5 face-to-face sessions: an initial session where they received their laptop and 4 monthly sessions where we collected the data and analyzed the computer. Participants were invited to book their appointments via an on-line calendar system hosted on our server. To encourage subjects to remain in the study, we paid them for each session attended. If they completed all sessions, a bonus was paid out; in total, if a subject attended all sessions they would receive a sum equivalent to the cost of the laptop, along with a small additional compensation.

**Initial session** The intent of this short session was to obtain each user’s informed consent and provide them with their laptop. Each user had to read and sign the informed consent form to confirm their participation in the study. Thereafter, the laptop was sold at a below retail-market price to the users, with laptops staying in users’ possession after the study. This option was chosen for legal reasons and to foster user ownership of their computer, in the hope of reducing experiment bias in user behaviour. The only restrictions imposed were that they were not allowed to do the following during the study: i) format the hard drive, ii) install another operating system, iii) delete our tools and the data collected, iv) install another AV product, and v) create a new disk partition. In addition, users were asked to answer an initial questionnaire to collect general information for their profile, such as gender, age group, status (worker, student, unemployed), field of expertise (computer science, natural science, art and humanities) and self-reported level of computer expertise.

**Monthly sessions** During the monthly sessions, users answered an online questionnaire. The aim of this questionnaire was to assess user experience and opinion of the AV product, gain insights about how the computer was used, determine their level of security awareness and their reported due diligence exerted to secure their computers. Meanwhile, statistical data compiled by the scripts were collected on the computer by the experimenter. The computer was also analyzed following a strict, fixed protocol, to look for malware missed by the AV product. The following diagnostic tools were used:

- HijackThis: gives the list of auto-loading programs and services, BHOs, plugins, tool-bars, etc.;

- ProcessExplorer: shows the list of active processes;
- Autoruns: gives the complete list of programs configured to run during system bootup or login;
- Sigcheck: shows file version number, timestamp information, and digital signature details, including certificate chains;
- SpyBHORemover: gives the list of installed BHOs and classifies them in 4 categories (dangerous, suspicious, safe, unrated);
- SpyDLLRemover: gives the list of loaded DLLs and classifies them in 3 categories (dangerous, safe, unrated);
- Whatchanged: scans for modified files and registry entries;
- Winprefetchview: reads prefetch files and displays information stored in them.

We classified each element in 4 categories (safe, dangerous, suspicious, unrated) using external on-line resources, such as *www.systemlookup.com*, *www.processlibrary.com*, VirusTotal (Virus Total, 2013), and Anubis (International Secure Systems Lab, 2013). Computers with files identified as dangerous or suspicious were suspected to be infected, and any unrated files were subject to an in-depth investigation to see if they had malicious purposes. If the AV product detected malware over the course of the month, or if our diagnostic tools indicated that the laptop was infected or suspected to be, users were asked to answer an additional questionnaire. This specific questionnaire collected more information regarding the potential means and sources of the infection, and on any behavioural changes observed on the computer. Moreover, additional consent was requested from the users to collect specific data, such as the browser history, network traffic data from tshark log files, and the suspicious file(s). These data were collected to help us identify the vector and the source of the infection.

**Final session** The final session was similar to the other monthly sessions. However, users answered a post-experiment questionnaire about their overall experience in the study. This final survey helped us identify activities or mindsets that may have unduly affected the experimental results. We also requested that users keep their experiment data for an additional period of 3 months in the event we might need to perform more in-depth analysis of their computer. Finally, we provided procedures to stop the automatic collection of the data, delete the data and the tools we installed, and reinstall the operating system, if they wanted to do so.

## 5.5 Antivirus evaluation

To evaluate the AV product, we analysed the detections (blocked malware attacks) and the missed detections (successful malware attacks) occurring over the course of the experiment. Additionally, users' questionnaire responses were compiled to provide an overall picture of the AV's subjective performance.

### 5.5.1 Threats detected by antivirus

During the 4-month study, 380 suspicious files were detected on 19 different user machines by the AV product being evaluated. However, some of these files were detected multiple times on the same user machine. Removing these repetitions, we obtain a total of 95 unique detections. Figure 5.1 shows the frequency of unique detections. The minimum number of detections observed per user is 0, the maximum is 28, and the average number of detections per user is 1.19 ( $SD = 4.46$ ). Among those 95 unique detections, we were able to trace the source of infection and determine that 17 of these propagated through portable storage devices (USB key or external hard drive).

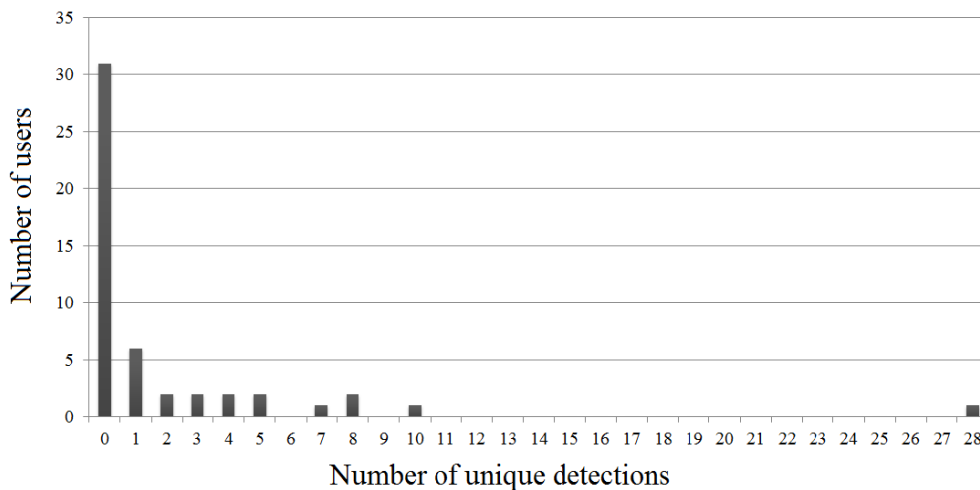


Figure 5.1 Frequency histogram of unique detections

In terms of overall virulence, 38% of the users were exposed to computer threats over a period of 4 months. More importantly, however, these results indicate that, if they are representative of the whole Internet population, 1 out of 3 newly installed machines would have been infected within 4 months if they had not had an AV installed. This figure aligns with the Eurostat Annual Report Eurostat (2011) indicating that over a period of 12 months in 2010, 31% of users reported a virus infection on their home computers, while 84% of these users reported

having some kind of security software installed (e.g. AV, anti-spam, firewall). Regarding the evolution of detections over time, the level of monthly detections is quite stable, as shown in Figure 5.2.

Detections were classified based on information provided by the AV product. As illustrated in Figure 5.3, most detections were classified as trojans, while viruses and adware had a relatively weak representation. These results are somewhat similar to those reported for overall detections by other AV vendors for the same period. For example, the 2011 Annual Report from Panda Security Panda Security Labs (2011) indicates that trojans account for most detections with a ratio of 73%, while worm, virus, adware and other have respective ratios of 8%, 14%, 3% and 2%. The differences with our results could be partially attributed to differences in the classification methods. For example, a file might be classified as a trojan by the AV product being evaluated and as a virus by another product. Furthermore, statistical error could be significant since our results are only based on 95 samples, while Panda Security has access to thousands of different samples and a user base of several millions users.

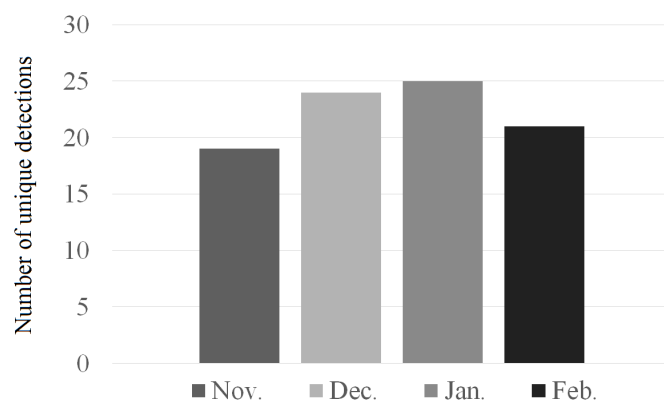


Figure 5.2 Unique malware detections per month

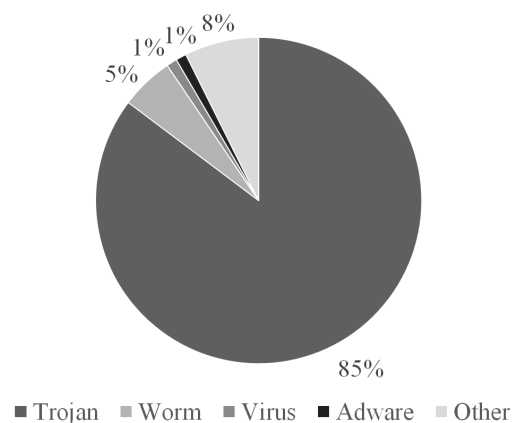


Figure 5.3 Malware detections by type

### 5.5.2 Missed threats

The process of identification and classification of missed detections was based on user reporting of suspicious machine behaviour, monthly analysis of logs from the diagnostic tools, and results of automated queries to on-line sources with respect to processes and files found on the machine, and start-up programs (obtained automatically by scripts that we wrote).

Overall, 20 possible infections were detected on 10 different machines. The most useful diag-

nostic tool was HijackThis, which was involved in identifying 18 of the suspected infections. SpyBHORemover uncovered one additional infection. The last suspected infection was reported by the user, who contacted the project manager when he suspected that his machine had been infected. Except for the user-reported suspected infection, all suspicious files were captured during the monthly visits. While the logs show the location and filename, the file could not be retrieved as it seems that the suspected malware uninstalled itself between the time the user called in and the following lab visit.

All captured files (19 out of 20) were later scanned with the evaluated AV product to see if they would be detected *a posteriori*. Even several months after the end of the experiment, none were detected by the AV product or identified as a potential threat. We scanned the captured files *a posteriori* with the VirusTotal service to compare the results obtained by several AV products and to compare these later results with those obtained a few months earlier. Additionally, we searched the Internet to find as much detail as we could for each of these 20 detections. From this analysis, we classified twelve samples as unwanted software, seven as adware, and one as rogueware, for a total of 20 missed threats.

The 12 detected unwanted software and 7 adware samples were either BHO or toolbars. In all cases, they were unknowingly installed by the users. Their effects included changing the web browser home page, redirecting web searches, or displaying advertisements. However, it was unclear if the adware samples were indeed malicious, in that they show additional behaviour (e.g. theft of personal/private information) that might have further consequences for the user. The last sample was identified as rogueware —a software that pretends to be an AV program but does not provide any security. As previously mentioned, the corresponding user informed the project manager that his laptop was probably infected. It turned out that the laptop was infected with a fake AV product (AV Security Scanner). Warning windows were regularly appearing to inform the user that harmful software was on his computer and every application started was killed except for web browsers. To get rid of these infections, the user was invited by the rogueware to register and provide his contact and payment information. At that moment, the user suspected that he may be infected and contacted the project manager. Since the files disappeared from the computer before it was brought in for inspection, it was not possible for us to verify if the AV product could detect this threat *a posteriori*.

Overall, 20 missed threats were detected on 10 machines, which represents 20% of users. If we consider only missed malware, i.e. the seven adware and the rogueware, 12% of users got infected. One point of comparison is the 2009 SurfRight report on real-world malware statistics SurfRight (2009). Over a period of 55 days, 107,435 users scanned their machine

with the Scan Cloud product. Among those, Scan Cloud found that 32% of protected machines were infected, compared to 46% of unprotected machines. In comparison with our 20% and 38% ratio, it would appear that our users were less at risk than those using SurfRight’s Scan Cloud. One possible explanation is simply that one of the motivations for using such a product is that users already suspect that their machines are infected, therefore resulting in an important self-selection bias. In all cases, direct comparison with our study is difficult given the fact that the time-period and the definition and classification methods for threats are quite different.

### 5.5.3 Antivirus efficacy

The *efficacy* of the AV product (AE) is a function of the number of actual threats detected, i.e. the *true-positives* (TP), and the number of threats missed, i.e. the *false negatives* (FN).

$$AE = \frac{TP}{TP + FN} \quad (5.1)$$

If we add the 20 threats that were not detected by the AV (FN) to the 95 unique detections (TP), the AV has been exposed to a total of 115 threats.

$$AE \approx 0.8261 \quad (5.2)$$

Therefore, the AV product provided an efficacy of 83% . More specifically, this result represents the sensitivity of the AV to properly identify threats, including malware, and potentially unwanted software. If we only consider missed malware, i.e. the seven adware and the rogue-ware, the efficacy raises to 92%. In comparison, the test performed by PC Security Labs PC Security Labs (2013) for the same period reported an efficacy of 99% for Trend Micro, and AV-Comparatives AV Comparatives (2013) reported an efficacy of 98% for the same product and period. These differences in performance between our test and the commercial tests suggest that AV protection differs between real-life conditions and controlled conditions. In other words, AV *field efficacy*, i.e. how the AV performs in actual use, is lower than AV efficacy from in-lab evaluations where testers have greater control of the testing environment.

### 5.5.4 User experience

We evaluated user experience with the AV product through monthly surveys. We assessed their opinion regarding the level of interference, the information provided by the AV, the perceived level of protection provided, and their attitude toward the AV product. In addition,

every time malware was detected by the AV, or the laptop was suspected to be infected by our diagnostic tools, we also collected insights on how users interact with the AV and potential computer threats while using their system.

**Experience with the AV** Descriptive statistics relating to user perceptions for the 4 months (M1, M2, M3, M4) of the study are presented in Table 5.1. We computed the relative frequency of responses for questions Q1 to Q4, and the arithmetic mean (AM) and the standard deviation (SD) for questions Q5 to Q8.

Overall, 1/3 of users mentioned that the level of interaction (Q1) required by the AV was not enough and 2/3 judged that the required level was adequate. Only a few users found the level of interaction too high. Those findings are also confirmed by the monthly results on the level of interference (Q5), which ranged from 3.0/10 to 3.4/10, where 1 meant no interference and 10 meant high interference. It is worth mentioning that the AV evaluated (i.e. Trend Micro OfficeScan) was a business product that was configured to be silent. Pop-up windows would only appear in the case of a detection, which could explain why one third of the users found that the level of interaction was not enough. One potential explanation could be that for those users, interaction with the AV provides reassurance that they are protected (Furnell, 2010).



Table 5.1 User experience and opinion per month

	M1	M2	M3	M4
Q1 Level of interaction required by the AV				
Too frequent	2%	4%	4%	2%
Adequate	68%	59%	55%	60%
Not enough	30%	37%	41%	38%
Q2 Amount of information provided by the AV				
Too much	2%	2%	0%	0%
Adequate	47%	51%	49%	48%
Not enough	51%	47%	51%	52%
Q3 Response to a pop-up window from the AV				
I read it and follow its suggestions	60%	59%	65%	68%
I read it but don't follow its suggestions	11%	6%	10%	4%
I close it without reading it	6%	16%	4%	12%
Other (specify)	23%	19%	21%	16%
Q4 Feeling when an AV's pop-up window appears				
Comforted to know that the AV is working	60%	70%	63%	72%
Annoyed that the AV is interrupting me	19%	4%	10%	4%
Don't notice	15%	10%	10%	8%
Other (specify)	6%	16%	17%	16%
	AM (SD)	AM (SD)	AM (SD)	AM (SD)
Q5 Level of interference (1 to 10)	3.4 (2.8)	3.2 (2.7)	3.0 (2.3)	3.4 (2.6)
Q6 Level of usefulness of the information (1 to 10)	5.7 (2.2)	6.1 (2.5)	6.0 (2.4)	6.3 (2.4)
Q7 Level of protection (1 to 10)	7.8 (2.0)	7.7 (1.9)	7.5 (2.1)	7.7 (1.8)
Q8 Level of understanding of the information (1 to 10)	6.8 (2.2)	6.8 (2.3)	7 (2.5)	6.7 (2.6)

When evaluating the information provided by the AV (Q2), half the users found that the level of information was adequate and half found it was not enough. The product was configured to only give the name and the path of the file detected, the action from the AV, and generic information on the family. For half the users, this information was not sufficient, meaning that a minimalist design might not be appropriate for all users. The level of usefulness (Q6) of the information provided ranged from an average of 5.7/10 to 6.3/10, where 1 meant useless and 10 meant very useful. We also evaluated if the information provided by the AV was presented in a manner that could be easily understood by users (Q8). Monthly average results ranged from 6.8/10 to 7.0/10, where 1 meant difficult to understand and 10 meant easy to understand.

We also asked users how they felt (Q4) and reacted (Q3) when they saw a pop-up window from the AV. More than 2/3 of users said they feel comforted to know that the AV is working. The last third was either annoyed that the AV was interrupting them or did not notice any pop-up. Most users who answered “Other” mentioned that they did not get any pop-ups from the AV. Examples of other answers included: “I don’t want to see the pop-up when I am watching a movie. but other times, I don’t care”, “I don’t understand what happens”, “It doesn’t bother me”, or “I feel annoyed because there is a virus on my computer”.

Regarding how users reacted when they saw a pop-up window (Q3), almost 2/3 reported that they read and followed the suggestions of the AV. Over the 4 months, between 4% and 11% of the users said they read the pop-ups but ignored the suggestions. And between 4% and 16% of users said they closed the pop-ups without reading them. Most participants who answered “Other” mentioned that they did not experience any pop-up from the AV. Some “Other” answers included: “I read it as well as the suggestions but I take the action I want”, “I read it and sometimes I follow its suggestions”, “It depends”, “I read it quickly and if it’s important I follow the suggestions, if it’s not, I close it”, and “I ask someone else to take care of it”.

The perceived level of protection (Q7) provided by the AV was also evaluated over the study. The monthly averages ranged from 7.5/10 to 7.8/10, where 1 meant very low protection and 10 meant high protection. We computed the average level of perceived protection provided over the 4 months for users that had at least one detection from the AV, and for users that had no detection. Users that experienced a detection by the AV over the study had an average level of perceived protection of 7.38/10 ( $SD = 1.66$ ), and users that had no detection had an average of 7.83/10 ( $SD = 1.48$ ). A Mann-Whitney U-test was conducted and no statistically significant difference was found between users with and without detections;  $U = 243.00$ , two-tailed exact p-value = 0.31. We also conducted a Mann-Whitney U-test to

investigate if there was a significant difference between users that had at least one missed threat by the AV ( $A.M. = 7.07, SD = 1.37$ ), and users that had no infection ( $A.M. = 7.81, SD = 1.57$ ). Results of the test indicated that there is no significant difference;  $U = 133.50$ , two-tailed exact p-value = 0.11. Although no significant difference was found, users that experienced no detection or missed threats over the study reported a marginally higher level of perceived protection.

**Experience with computer threats** Additional information on user experience with potential computer threats were collected during the monthly sessions when the AV detected malware or when we suspected the laptop to be infected. All concerned users agreed to answer the additional questionnaire and provide more specific data about the system’s activity.

As part of the questionnaire, users were asked to report any strange computer behaviour they might have experienced over the last month. Of the 40 reports, 22 users said they had not observed strange behaviours, 2 said they did not know, and 16 answered yes. Examples of strange behaviour included annoying pop-ups, music starting to play, new web browser home page, web search redirections, and changes in computer performance (e.g. crashes or slowdowns). Among the reports that were related to missed threats and not threats detected by the AV, 7 users said they did not observe strange behaviour, and 8 said they did. While half the users observed behaviour that are known to be warning signs of malware infection, the other half did not notice anything abnormal on their computer even though they were infected with some form of computer threat that was missed by the AV.

We also asked users if they remember receiving any security-related messages from the system or the AV. Interestingly, only half of users answered yes. Among those, 6 said they felt comforted, 4 mentioned that they felt annoyed by the interruption, 3 were confused, 6 were worried about the security of their computer, and 1 answered “Other” (“I was concerned about the computer’s security, but I would like to proceed on that”). In addition, we asked users to report what they were doing when the message appeared. Only 5 users said they did not remember. The other 15 responses included: visiting web sites ( $N=3$ ), downloading software/files from the Internet ( $N=9$ ), using a portable storage device ( $N=2$ ), and watching a movie ( $N=1$ ). While those comments are not sufficient to establish the exact transmission vectors used, they suggest potential user involvement in the infection process, whether blocked by the AV or successful.

## 5.6 User Profiling and Behaviour

We examined whether user demographics, characteristics, and certain types of user behaviour led to a higher probability of malware attack. We first divided users in two groups. The first group contains *high-risk* users, which are those who experienced at least one malware attack, whether blocked or successful, and the second group contains *low-risk* users who had no malware attack during the experiment. Table 5.2 shows the user distribution between the total sample, the *high-risk* group, and *low-risk* group, based on user characteristics and demographic factors.

The risk analysis was determined based on the calculation of the odd ratio (OR) —a measure of the degree of association between a risk (or protective) factor and an outcome. It represents the ratio between the probability that an outcome will occur in a group exposed to a factor of interest and a reference group that is not exposed. Given that A is the number of individuals in the exposed group who developed the outcome, B is the number of individuals in the exposed group who did not develop the outcome, C is the number of individuals in the reference group who developed the outcome, and D is the number of individuals in the reference group who did not develop the outcome, the OR can be calculated as follows:

$$OR = \frac{A * D}{B * C} \quad (5.3)$$

An OR larger than 1 indicates that the factor of interest is a risk factor. An OR smaller than 1 means that the exposure is a protective factor. And if the OR equals 1, the outcome is equally likely in both groups. The confidence interval (CI) in which the true value of the OR is likely to be must also be taken into account when interpreting the OR. Hence, if 1 is included in the CI, nothing can be said on the association between the factor and the outcome.

### 5.6.1 Characteristics and demographic factors

Risk analysis through OR was performed to assess if particular user characteristics and demographics increase the odds of malware attack. Malware attack was used as the outcome, indicated by either 1 or 0, depending on whether the user experienced any malware attack during the experiment. The factors of interest were gender, age, status, field of expertise, and self-reported level of computer expertise. Female, 18-24 age group, unemployed,

Table 5.2 Proportion of users for each factor

Factor		Total sample (N=50 users)	High-risk group (N=23 users)	Low-risk group (N=27 users)
Gender	Male	60%	61%	59%
	Female	40%	39%	41%
Age	18-24	38%	35%	41%
	25-40	46%	61%	33%
	41+	16%	4%	26%
Employment status	Student	64%	70%	59%
	Worker	30%	26%	33%
	Unemployed	6%	4%	8%
Field of expertise	Computer Science	26%	22%	30%
	Natural Science	52%	48%	56%
	Arts/Humanities	22%	30%	14%
Computer expertise	High	18%	30%	7%
	Low	82%	70%	93%

arts/humanities, and low self-reported level of computer expertise were used as the reference groups for gender, age, status, field of expertise, and self-reported level of computer expertise respectively. Results of the analysis are summarised in Table 5.3. For each factor, we computed the OR, the 95% CI, and the p-value as a measure of statistical significance. For the purpose of our analysis, items marked with \* were considered as statistically significant at  $p\text{-value} < 0.05$ .

Table 5.3 Odds ratio of user characteristics and demographic factors

Factor		OR	(95% CI)	p-value
Gender	Male vs. Female	1.06944	(0.34339-3.33061)	0.90778
Age	25-40 vs. 18-24	2.13889	(0.62071-7.37039)	0.02934 *
	41+ vs. 18-24	0.19643	(0.01999-1.92938)	0.07169
Employment status	Worker vs. Unemployed	1.33333	(0.09772-18.19174)	0.94323
	Student vs. Unemployed	2.00000	(0.16442-24.32783)	0.46665
Field of expertise	Computer Sci. vs. Arts/Humanities	0.72917	(0.15303-3.47431)	0.55542
	Natural Sci. vs. Arts/Humanities	1.16667	(0.30169-4.51161)	0.58426
Computer expertise	High vs. Low	5.46875	(1.00696-29.70058)	0.04907 *

**Gender** The total sample included 30 males and 20 females which gives a proportion of 60% and 40% respectively. Table 5.2 shows that the gender distribution among the 23 *high-risk* users is very similar to the total sample, indicating that gender may not be a significant risk factor for malware attack. This was supported by the statistical analysis where no

significant difference between males and females (Table 5.3) was found with respect to the risk of malware attack.

In comparison with previous studies that investigated the effect of gender, six out of eight studies reported a significant gender effect. Some researchers (Yen *et al.*, 2014; Reyns, 2013; Lalonde Lévesque *et al.*, 2017) found that males were more at risk than females, and others (Bossler et Holt, 2009; Jagatic *et al.*, 2007; Oliveira *et al.*, 2017) found that females were more susceptible to computer threats than males. While our results are in line with the studies that reported no significant effect (Grimes *et al.*, 2007; Kumaraguru *et al.*, 2009), direct comparison is not possible, we studied malware attacks while Grimes *et al.* (Grimes *et al.*, 2007) used (self-reported) reception of spam and Kumaraguru *et al.* (Kumaraguru *et al.*, 2009) investigated phishing susceptibility. When looking only at studies that focused on malware attacks Yen *et al.* (2014); Lalonde Lévesque *et al.* (2017), males were found to be more at risk of encountering malware than females. This discrepancy with our results could be attributed to differences in study design, target population, and sample size; Yen *et al.* (2014) studied malware encounters of corporate users within a large enterprise, and Lalonde Lévesque *et al.* (2017) based their study on malware encounters of millions of Windows users. Although prior work (Yen *et al.*, 2014; Lalonde Lévesque *et al.*, 2017) suggests that gender is a significant correlate of malware attack, further studies should be conducted to validate the direction of the aforementioned correlation, if any.

**Age** We divided users into three age groups as evenly as possible (although we note that the older age group has fewer users due to our sample). Table 5.2 shows that the proportion of 18 to 24 year-old in the *high-risk* group is almost the same as for the total sample. For those 25 to 40, the proportion in the *high-risk* group (61%) is higher than for the total sample (46%), which could suggest that this age group is more susceptible to malware attack. And for the 41+ age group, we observe a decrease of 12% in the proportion between the total sample (16%) and the *high-risk* group (4%). Results from the analysis (Table 5.3) revealed a significant difference between the 25-40 age group and the reference group (18-24). However, as the value 1 is included in the 95% CI, nothing can be said on the nature of the association, that is whether it is a risk factor or a protective factor.

Similarly to most prior work that included the effect of age, our statistical results mildly suggest it may be a contributing factor associated with the risk of malware attack. In comparison, some researchers found younger users to be more susceptible to phishing (Sheng *et al.*, 2010; Kumaraguru *et al.*, 2009; Jagatic *et al.*, 2007) and malware attacks (Ngo et Paternoster, 2011; Lalonde Lévesque *et al.*, 2017), while Reyns (2013) found older users to be significantly more at risk of (self-reported) identity theft. Bossler et Holt (2009) and Grimes

*et al.* (2007) reported no significant age effect on (self-reported) data loss from malware infection and (self-reported) reception of spam respectively. These discrepancies can be explained first because the experimental methods are quite different: some studies involved surveys of users where susceptibility levels are evaluated through user self-declarations of previous incidents, and not from actual observation. Second, these results are not (all) specific to malware attacks. Finally, the granularity of the age data recorded is different so it is hard to precisely compare these discrepancies, especially since the age distributions are quite different. In any case, what is clear is that none of these studies, including ours, can be used to make categorical statements about risk of malware attack and age. Large scale studies based on alternate data sources, other time frames, and different analysis methods will be required to settle the issue of age as a contributing factor for malware attack.

**Employment status** Users were classified in three self-declared categories: student, worker, or unemployed. Table 5.2 indicates that the proportion between the total sample and the *high-risk* group is quite similar for each category, suggesting that employment status may not be a contributing factor of malware attacks. This was confirmed by the risk analysis (see Table 5.3) where no statistically significant difference is shown between the different categories.

In contrast, prior work that studied the effect of employment status found employed users to be at higher risk of (self-reported) data loss from malware infection (Bossler et Holt, 2009) and (self-reported) identity theft (Reyns, 2013). Given that unemployed users represented only 6% of our study, it is possible that our sample was simply too small to observe any significant difference.

**Field of expertise** We recruited users based on their field of work or study in order to have a heterogeneous sample. As shown in Table 5.2, 26.5% of users were self-identified as being in computer science, 47% in natural science, and 26.5% in arts and humanities. Although the table suggests that those in the arts/humanities might be slightly more at risk, results of the risk analysis (Table 5.3) show no statistically significant effect for the field of expertise.

To the opposite, Yen *et al.* (2014) found that job types have a significant impact on the risk of malware encounters; jobs from the top of the enterprise organizational tree and jobs requiring higher technical expertise had a greater likelihood of malware encounter. Similarly, Thonnard *et al.* (2014) identified directors, high-level executives, and personal assistants to be at higher risk of targeted attacks compared to other jobs. Finally, Lee (2012) found that some areas of work are associated with increased risk of being subjected to targeted

phishing attacks, suggesting that it is the area of expertise that leads users to be of interest to attackers. Although prior studies found that the field of expertise may be a contributing factor, direct comparison with our results is not possible as we studied *home users* while they focused on non home-user domains (e.g. industry, government, academia).

**Computer expertise** We assessed computer expertise by asking users about their proficiency with certain technical tasks. Users were considered to have a high self-reported level of computer expertise if they had previously completed all of the following tasks: configured a home network, created a web page, and installed or re-installed an operating system on a computer. Overall, 18% of users were classified as self-reported computer “experts” for the purposes of our analysis. As observed in Table 5.2, those with high expertise were nearly twice as likely to be in the *high-risk* group when compared to the total sample. This may indicate that a high level of expertise increases the risk of malware attack, which was confirmed by the statistical analysis. More specifically, users with high self-reported level of computer expertise were found to be 5.47 times more likely to experience malware attack than users with low expertise.

Although our results are somewhat counterintuitive, they are consistent with the work of Ovelgönne *et al.* (2017) and Yen Yen *et al.* (2014). Ovelgönne *et al.* (2017) identified software developers to be more prone to malware attack, and Yen *et al.* (2014) found people with technical expertise to be more at-risk of encountering malware. In opposition, Onarlioglu *et al.* (2012) found *computer security expertise* to be a protective factor. A possible explanation is that self-reported expert users are more at risk of malware attack because they know just enough to get themselves into trouble. For example, they may have a false sense of self-confidence that leads them to engage in more risky behaviours. Another potential explanation could be that users with high computer expertise have a high risk-seeking profile, which lead them to engage in risky behaviours. One last explanation could be that expert users are *heavy* computer users (they spend more time online, they download more applications from the Internet, etc.), which contributes, intentionally or not, to increase their odds of getting exposed to malware.

**Summary of user characteristics and demographic factors** In summary, we found little evidence linking user demographics and characteristics to increased risk of malware attack. Gender, student/employment status and field of expertise showed no statistically significant differences. However, we did find partial support linking age and self-reported level of computer expertise to the risk of malware attack.



### 5.6.2 Behavioural factors

To assess if specific user behaviour led to a higher risk of malware attacks, we focused our analysis on the following factors: system activity, application installs, network usage, web browser usage, web pages visited, files downloaded, peer-to-peer (P2P) activity, and level of vigilance. Data was collected through scripts on the computer and self-reported questionnaires. Using a similar approach to that described in Section 5.6.1, we conducted a risk analysis based on the calculation of the OR. In the case of continuous variables, the OR is interpreted in terms of each unit increase on the variable; for each increase by one unit, the odds of the outcome is multiplied by the OR. Table 5.4 summarises the statistical results; items marked with \* were considered statistically significant at  $p\text{-value} < 0.05$ .

Table 5.4 Odds ratio of behavioural factors

Factor		OR	(95% CI)	p-value
System activity		1.00047	(0.99972-1.00121)	0.22066
Applications installed		1.00678	(0.99449-1.01922)	0.28083
Outdated applications		1.03763	(0.75098-1.43369)	0.82282
Connection time		1.00369	(1.00044-1.00697)	0.02618 *
Hosts contacted		1.00002	(1.00000-1.00005)	0.04969 *
Default web browser	Firefox vs. IE	1.83333	(0.39238-8.57580)	0.74626
	Chrome vs. IE	5.10714	(1.17708-22.15903)	0.03005 *
Web pages visited		1.00007	(1.00002-1.00013)	0.00697 *
Files downloaded		1.00007	(0.99956-1.00196)	0.21369
P2P activity	Yes vs. No	13.63636	(2.60209-71.46171)	0.00199 *

**System activity** The activity of the system was measured by scripts using the number of hours per day the laptop was on. To study its impact on the risk of malware attack, we computed the total number of hours the laptop was on for the entire duration of the study. The total system usage ranged from 109 hours to 2,882 hours, with an average of 1,629 hours ( $SD = 778$ ). When comparing groups, *high-risk* and *low-risk* users had their laptop on for an average of 1,793 hours ( $SD = 656$ ) and 1,522 hours ( $SD = 863$ ) respectively. Results from the analysis in Table 5.4 show no significant relationship between the system activity and the risk of malware attack. Hence, our analysis suggest that the system activity —as measured by the number of hours the system was on— does not seem to be a significant factor for malware attack.

**Application installs** We monitored using scripts the daily number of applications installed by each user. To assess the potential effect on the risk of malware attack, we computed for

each user the total number of applications installed over the 4 months. Users installed between 2 and 177 applications, with an average of 70 (SD = 47) applications. The *high-risk* group installed on average 75 (SD = 46) applications, while the *low-risk* group installed 61 (SD = 47) applications on average. However, this difference was not found to be significant from the risk analysis (see Table 5.4). In contrast, Ovelgönne *et al.* (2017) found a significant positive correlation between the number of binaries installed, and the number of attempted attacks per host. For comparison, we computed the correlation between the number of unique malware attacks and the number of applications installed. The Gamma statistic, a non-parametric correlation coefficient, was used because our data on malware attacks contains many tied observations. Similarly to Ovelgönne *et al.*, we found a weak significant positive relationship ( $G=0.24$ ,  $p\text{-value}=0.04$ ,  $N=50$ ) between the number of applications installed and the number of malware attacks. This seems plausible as installing many applications can contribute to increased probability of being exposed to malware, either by a malicious application, or by a legitimate application that may install unwanted software.

We also investigated the type of applications that were installed. Users were asked through the monthly survey what type of applications they installed the most (see Table 5.5), and what type of applications was installed by other people (see Table 5.7). Table 5.5 shows that the majority of applications installed over the study were not reported as games. Moreover, there does not seem to be major differences in the type of applications between users in the *high-risk* group and in the *low-risk* group, given the high level categorization used in the questionnaire. From Table 5.6, we see that the majority of users reported that no one besides them has installed applications on their computer. When comparing *high-risk* and *low-risk* groups, *high-risk* users more frequently reported that others had installed applications on their computer, which could suggest that *high-risk* users are more likely to let other people use their computer.

Table 5.5 Type of applications installed per month

	High-risk group				Low-risk group			
	M1	M2	M3	M4	M1	M2	M3	M4
Most of the applications are games	5%	4%	0%	0%	0%	11%	8%	4%
Most of the applications are not games	90%	73%	61%	70%	88%	63%	60%	59%
No application was installed	0%	9%	30%	26%	8%	22%	28%	37%
Other	5%	14%	9%	4%	4%	4%	4%	0%

The survey also asked the most frequent and the second most frequent type of applications used. From Table 5.7, we can see that between 82% and 96% of users used a web browser most frequently. Half of participant reported that the Microsoft Office suite was second most

Table 5.6 Type of applications installed by others per month

	High-risk group				Low-risk group			
	M1	M2	M3	M4	M1	M2	M3	M4
Most of the applications are games	14%	5%	5%	4%	0%	8%	4%	0%
Most of the applications are not games	19%	18%	17%	22%	15%	18%	8%	26%
I don't know	0%	5%	17%	9%	8%	4%	11%	0%
No one besides me has installed applications	62%	72%	61%	61%	73%	70%	77%	70%
Other	5%	0%	0%	4%	4%	0%	0%	4%

frequently used (Table 5.8), followed by web browser and other. Comparison between the *high-risk* and the *low-risk* groups does not suggest major differences; they both reported web browser and Office suite as their most and second most frequently used applications.

Table 5.7 Most frequently used applications per month

	High-risk group				Low-risk group			
	M1	M2	M3	M4	M1	M2	M3	M4
Web browser	86%	82%	91%	83%	88%	92%	96%	88%
Office Suite	0%	5%	9%	13%	8%	4%	0%	4%
Mail application	0%	0%	0%	0%	0%	0%	4%	4%
Games	0%	0%	0%	0%	0%	0%	0%	4%
Other	14%	13%	0%	4%	4%	4%	0%	0%

Table 5.8 Second most frequently used applications per month

	High-risk group				Low-risk group			
	M1	M2	M3	M4	M1	M2	M3	M4
Web browser	14%	23%	17%	26%	8%	11%	8%	15%
Office Suite	43%	45%	61%	52%	46%	48%	54%	63%
Mail application	10%	9%	4%	0%	19%	15%	12%	7%
Games	10%	5%	4%	4%	15%	22%	19%	11%
Other	23%	18%	14%	18%	12%	4%	7%	4%

In addition, we also investigated the number of applications for which updates were available, as outdated applications may increase the odds of malware infection. We computed the 4-month average number of outdated applications per user. Overall, users had on average between 3 and 11 outdated applications, with a mean of 7 (SD = 2) outdated applications. When looking at the *high-risk* and the *low-risk* group, both had on average 7 outdated applications. Based on the risk analysis (Table 5.4), the average number of outdated applications does not seem to be a significant risk factor.

**Network usage** User network activity was evaluated in terms of time spent online, number of different hosts contacted, and reported primary connection location. To assess the relationship between the time online and the risk of malware attack, we computed using scripts the total number of hours each laptop was connected to the Internet for the entire duration of the study. The connection time varied from 11 hours to 992 hours, with an average of 242 hours per user ( $SD = 229$ ). *High-risk* users were connected on average 328 hours ( $SD = 273$ ), while *low-risk* users were connected on average 169 hours ( $SD = 155$ ). Results from the risk analysis in Table 5.4 show a weak significant positive association between the connection time and the risk of malware attack ( $OR = 1.00369$ ). For each 100 hours connected online, the odds of malware attack increase by 1.048 ( $1.00369^{100}$ ).

The daily number of different hosts contacted by the laptop was also collected over the 4-month period. For each user, we computed using scripts the total number of hosts contacted during the entire study. Users contacted between 18 and 1,508,833 hosts during the 4 months, with an average of 60,433 hosts per user ( $SD = 211,244$ ). *High-risk* users contacted a higher number of hosts during the study than *low-risk* users; they respectively contacted on average 107,268 ( $SD = 309,065$ ) and 20,536 ( $SD = 21,867$ ) hosts. From the risk analysis in Table 5.4, there is a weak significant association between the number of hosts contacted and the risk of malware attack. However, as the value 1 is included in the CI, nothing can be said about the nature of the association.

Table 5.9 Primary location from which the laptop was connected to the Internet per month

	High-risk group				Low-risk group			
	M1	M2	M3	M4	M1	M2	M3	M4
Home	81%	82%	78%	70%	81%	78%	85%	86%
University campus	9%	18%	18%	26%	11%	15%	4%	7%
Work	0%	0%	4%	0%	8%	7%	11%	7%
Coffee shop	5%	0%	0%	4%	0%	0%	0%	0%
Other	5%	0%	0%	0%	0%	0%	0%	0%

We also asked users through the monthly survey the primary location from which the laptop was connected to the Internet. When looking at the results in Table 5.9, between 70% and 86% of users answered home as their primary connection location, followed by university campus (4%-26%), and work (4%-11%). Both *high-risk* and *low-risk* groups reported home as their primary location, suggesting that primary location may not be a contributing factor to malware attack.

**Web browser usage** Each month, users were asked which web browser was installed, which one they used most and if they have changed the default security and privacy settings

of their browsers. For each factor, except for the question related to the default settings, we also collected real data usage from scripts during each monthly meetings. We therefore prioritized, when possible, real data usage for our analysis instead of self-reported data obtained through surveys.

Table 5.10 presents the proportion of users that installed each web browser during the study, and Table 5.11 summarises the proportion of users who used each web browser. An increase of 17% is observed between the total sample and the *high-risk* group for Chrome. In contrast, the proportion decreases for Firefox and IE. When looking at the risk analysis in Table 5.4, Chrome was identified as a significant risk factor. Users with Chrome as their default browser were found to be 5.11 times more likely to experience malware attacks than users of IE. While these results suggest that having Chrome as a default web browser is a significant correlate of malware attacks, they do not imply that using Chrome is in itself a contributing risk factor. Possible explanations could be differences in browser’s architecture or threats landscape. Another potential explanation could be differences in users. For example, Chrome users might have a high risk-seeking profile, or be *heavier* computer users compared to IE users.

Table 5.10 Installed web browsers

	Total sample	High-risk group
IE	78%	70%
Firefox	58%	65%
Chrome	66%	78%

Table 5.11 Most frequently used web browser

	Total sample	High-risk group
IE	30%	17%
Firefox	30%	26%
Chrome	40%	57%

As many web browser offer advanced security and privacy settings, such as anti-phishing or anti-malware protection, we also investigated the effect of those changes on the risk of malware attack. Out of 50 users, only 4 changed the default security and privacy settings of their main browser. One disabled cookies for Chrome, another asked Chrome to remember all of his passwords, and the last one decided not to keep his IE temporary files. Since only a small proportion of users changed their default settings (see Table 5.12), we cannot draw any conclusion on the effect of those changes.

Table 5.12 Security and privacy default settings

	Total sample	High-risk group
Using default settings for all browsers	94%	96%
Made changes for Internet Explorer	2%	0%
Made changes for Firefox	0%	0%
Made changes for Chrome	4%	4%
Other	0%	0%

**Web pages visited** The number of web pages visited was also recorded for the entire duration of the study to evaluate the impact on the risk of malware attack. This factor was computed from the browser history using scripts and represents the total number of web pages visited by user. In total, users visited on average 18,531 (SD = 17,008) web pages. The *high-risk* group visited on average 26,624 (SD = 20,822) web pages while the *low-risk* group visited on average 11,637 (SD = 8,426) web pages. The risk analysis (see Table 5.4) reveals a weak positive association between the total number of web pages visited and the risk of malware attack (OR = 1.00007); for each 100 web pages visited, the odds of malware attacks increase by 1.007.

Our results confirm the general trend that the more a user surfs the web, the more likely he is to be exposed to computer threats. In comparison with previous work, Canali *et al.* (2014) also found that visiting many web pages increases the chance of visiting a malicious web page. In another study, Carlinet *et al.* (2008) reached a similar conclusion: heavy web activity, as measured by the web traffic, increases the likelihood of generating malicious traffic.

We further analysed if particular categories of web pages were more prone to be associated with malware attacks. To this end, each web page visited was classified using the Site Safety Center of Trend Micro Trend Micro (2012). Overall, 70 different categories of web pages were found. We performed a risk analysis based on the calculation of the OR using the 22 most popular categories (see Table 5.13). In total, 10 categories were found to be significant: streaming media/MP3, peer-to-peer, social networking, software downloads, email, personal network storage/file download servers, search engines/portals, games, entertainment, and computers/Internet. Among those, peer-to-peer, software downloads and personal network storage/file download servers were identified as the more risky. For each 100 web pages visited in these categories, the odds of malware attacks are multiplied respectively by 15.58, 10.60, and 7.56.

In comparison, Symantec Corporation (2012) identified the following 10 web site categories as the most *at-risk* of being malicious in 2011: blogs/web communications, hosting/personal web site, business/economy, shopping, education/reference, technology and Internet, entertainment and music, automobile, health and medicine, and pornography. Our findings are also similar to the results of Yen *et al.* that identified six web site categories as being associated with higher risk of encountering malware; chat, file transfer, freeware, social networks, and streaming. In another study, Canali *et al.* (2014) also identified that specific web site categories, such as pornography and adult content, were at higher risk of being malicious. Overall, those results suggest that *high-risk* categories are not limited to what common sense traditionally associates with higher risk, such as hacking and pornography.

Table 5.13 Odds ratio by web page categories

Factor	OR	OR <sup>100</sup>	(95% CI)	p-value
Streaming media/MP3	1.00168	1.18277	(1.00032-1.00305)	0.01582 *
Peer-to-peer	1.02784	15.57943	(1.00089-1.05551)	0.04276 *
Social networking	1.00018	1.01816	(1.00002-1.00034)	0.02440 *
Software downloads	1.02388	10.59024	(1.00642-1.04165)	0.00716 *
Pornography	1.00299	1.34791	(0.99702-1.00901)	0.32590
Email	1.00054	1.30234	(1.00005-1.00102)	0.02890 *
Personal network storage/file download servers	1.02044	7.56393	(1.00455-1.03658)	0.00697 *
News/media	1.00072	1.07463	(0.99984-1.00161)	0.11084
Shopping	1.00037	1.03769	(0.99959-1.00115)	0.35423
Chat/Instant messaging	1.00626	1.86647	(0.98623-1.02669)	0.54266
Search engines/portals	1.00056	1.05758	(1.00001-1.00110)	0.04485 *
Internet infrastructure	1.00788	2.19221	(0.99985-1.01598)	0.05454
Games	1.00736	2.08195	(1.00046-1.01431)	0.03642 *
Government/legal	1.00389	1.47439	(0.99933-1.00847)	0.09495
Entertainment	1.00409	1.50406	(1.00051-1.00767)	0.02508 *
Travel	1.00091	1.09522	(0.99906-1.00277)	0.33586
Blogs/web communications	1.00669	1.94794	(0.99861-1.01483)	0.10476
Financial services	0.99934	0.93611	(0.99745-1.00124)	0.49574
Business/economy	1.00104	1.10954	(0.99954-1.00254)	0.49574
Politics	0.99404	0.55003	(0.97494-1.01352)	0.54603
Computers/Internet	1.00127	1.13533	(1.00007-1.00246)	0.03688 *
Education	1.00055	1.05652	(0.99963-1.00147)	0.23837

**Files downloaded** For each user, we collected using scripts the number of files downloaded from the Internet over the study. During the 4 months, users downloaded between 19 and 3,341 files, with an average of 496 (SD = 588) files downloaded per user. Over the study, *high-risk* users (AM = 604, SD = 488) downloaded more files from the Internet than *low-risk* users (AM = 386, SD = 638). Though this may indicate that the volume of files downloaded from the Internet is a contributing factor of malware attacks, this factor was not found to be significant from our risk analysis (see Table 5.4).

We further investigated if specific types of files were associated with higher risk of malware attacks. We computed the OR for each file extension that had more than 100 files downloaded (see Table 5.14). Among the 9 types of files, only the extension *exe* was found to be a significant risk factor (OR = 1.06230). In comparison, Ovelgönne *et al.* (2017) also that found a positive association between the percentage of downloaded binaries from the web and the number of attempted malware attacks per host. Given that many malware are distributed via the Internet, it seems plausible that heavy downloading of executable files contributes to increased risk of being exposed to malware. The remaining question is whether

Table 5.14 Odds ratio by type of files downloaded

Factor	OR	(95% CI)	p-value
docx	1.14990	(0.82516-1.60244)	0.40939
rar	1.34096	(0.84383-2.13097)	0.21444
zip	1.03869	(0.98748-1.09256)	0.14115
pdf	1.00359	(0.99866-1.00855)	0.15379
exe	1.06230	(1.00846-1.11908)	0.02276 *
doc	0.95267	(0.83164-1.09134)	0.48439
ppt	1.15924	(0.93941-1.43051)	0.16839
jpg	1.01833	(0.99514-1.04207)	0.12224
gif	1.06268	(0.94728-1.19213)	0.29995

those executable files were downloaded by the users, or if they were silently downloaded from the Internet as a result of drive-by-download attacks.

**P2P activity** As part of the monthly survey, we asked users to report how often they have used peer-to-peer networks to download audio, video files or other software on the laptop. Overall, 14 users reported having engaged in peer-to-peer activities during the study. Among those, 12 were in the *high-risk* group and 2 in the *low-risk* group; suggesting P2P activity could be a risk factor. This was confirmed by the risk analysis in Table 5.4 where a strong significant association was identified between P2P activity and the risk of malware attack (OR = 13.63636). Users that reported engaging in P2P activity were found to be 13.64 times more likely to experience malware attack than users who did not. Our finding provides evidence that engagement in P2P activity might be a contributing risk factor of malware attack. This seems plausible as P2P networks are known to be a popular medium for spreading malware (Kalafut *et al.*, 2006).

**Level of vigilance** User level of vigilance was evaluated based on the level of security awareness and the measure of due diligence they exert to secure their laptops. Each month, users were required to report which of the following tasks they had previously completed: configured a firewall, secured a wireless network, and changed the default security and privacy settings of a web browser. Overall, 18% of users configured a firewall, 44% secured a wireless network, 44% changed the default security and privacy settings of a web browser, and 40% completed none of the above. As shown in Table 5.15, both groups reported similar expertise in computer security. Based on the number of tasks each user had previously completed, we computed a computer security score ranging from 0 to 3. From there, we performed a Mann-Whitney U-test and found no significant difference between both groups;  $U = 239.00$ , two-tailed exact p-value = 0.74. Though we found computer expertise to be a significant risk



factor, this was not the case for computer *security* expertise.

Table 5.15 Computer security expertise

	High-risk group	Low-risk group
Configured a firewall	17%	19%
Secured a wireless network	43%	44%
Changed the default security settings of a web browser	39%	48%
None of the above	22%	22%

We also evaluated through the monthly survey users' level of concern about the security of their laptop. The level of concern ranged from 1 to 10, where 1 meant low concern and 10 meant high concern. The 4-month average for the total sample was 7.27 (SD = 2.06). The *high-risk* group and the *low-risk* group reported similar level of concern; they respectively had an average level of concern of 7.14 (SD = 2.23) and 7.38 (SD = 1.94). In addition, we asked users to report on the tasks they performed, if any, to secure their laptop. Table 5.16 shows that a higher proportion of users in the *high-risk* group answered that they are concerned but they don't know what to do to secure their laptop from being compromised. In contrast, a higher proportion of users in the *low-risk* group said that they know what to do and they actively perform these tasks. The most common tasks mentioned were in order: avoid visiting dangerous and suspicious web sites, perform updates, avoid illegal downloading from the Internet, regularly scan computer, don't open suspicious files from the Internet, and perform risky actions in a virtual machine. Overall, we found no evidence linking the level of concern and the risk of malware attack. Rather, results suggest that being concerned is not sufficient if not combined with the adoption of safe computer behaviour.

Table 5.16 Concern about the computer's security per month

	High-risk group				Low-risk group			
	M1	M2	M3	M4	M1	M2	M3	M4
Typically not concerned	13%	13%	9%	13%	7%	11%	4%	11%
Concerned but don't know what to do	43%	52%	48%	48%	41%	30%	33%	30%
Know what to do but too busy	22%	17%	13%	22%	11%	4%	19%	15%
Know what to do and perform these tasks	13%	13%	26%	13%	37%	48%	41%	41%
Other	9%	4%	4%	4%	4%	7%	4%	4%

**Summary of user behaviour** We have identified six significant factors related to user behaviour: volume of network usage, number and types of web pages visited, default web browser, types of files downloaded from the Internet, and P2P activity. A high volume of network usage, as estimated by the time spent online and the number of hosts contacted, was

identified as a risk factor. Similarly, visiting many web pages as well as certain categories of web pages were found to be a contributing risk factor. We also found an association between the main web browser used and the risk of malware attack. Finally, downloading executable files from the Internet, and engaging in P2P activity were both found to increase the risk of malware attack.

## 5.7 Study limitations

The results we presented and discussed here are subject to certain limitations and potential bias that may threaten the internal and external validity of our study. Internal validity refers to the strength of the inferences from the study, that is the extent to which no other variables except the one we studied caused the results. While external validity refers to the ability to generalize the results to a more universal population.

First, the AV performance evaluation is limited to only 95 detected threats – a very small number compared to the numerous threats in the wild, especially considering that some of these may be false positives. As those threats were detected by an antivirus product, they depend on the efficacy of the later, which may lead to an underestimation of malware detections. In addition, the false negative number might also be underestimated because we cannot guarantee that our protocol caught all malware missed by the AV. In other words, we do not have absolute ground truth.

Second, even though we were able to identify several factors correlated with the risk of malware attacks, these factors in themselves are not sufficient to explain the *causal link* leading to malware infection. To this effect, a more detailed analysis of the collected data is required to determine the sources and means of infection for each of the 115 detected threats. Only then will we be able to determine which of these factors are causes of infection, and which are consequences of other factors that were not included in this study. Moreover, another limitation of our study is its susceptibility to confounding. Although we included in our analysis many variables that could influence the risk of malware attacks, and we fixed some of the external factors (same AV, laptop, OS, geographic area), we cannot guarantee that our results were not affected by other unknown extraneous variables that may confound the results. It would be interesting in future work to consider additional variables, such as culture, risk averseness, or risk propensity of users.

Another potential threat to the internal validity of our study is that users knew they were part of a computer security experiment. This knowledge might have caused them to alter their usage of their computer. We asked that question in the exit survey and 43 users claimed

that they did not modified their behaviour. Of the other 7 users, 2 admitted having modified their behaviour to fulfil experiment constraints (no OS reinstallation, creation of partitions, etc.), 2 others admitted voluntarily not performing potentially embarrassing activities on the computer, 1 mentioned refraining from visiting secure Internet banking sites, 1 admitted forcing himself to use the computer more frequently, and the last one explained that he controlled access to his computer to ensure being its only user. All in all, and considering that the usage statistics showed normal to high levels of computer and web activity, and that the computers were sold to and were to be kept by the subjects, we believe this potential experimental bias did not significantly affect our results.

One obvious limitation to the external validity of our study derives from our studied sample. First of all, subjects were located in the same geographic area. Second, their demographics (age and gender) and characteristics (status, field of expertise, computer expertise) distribution differ from that of the global Internet population. Third, we studied home users. Hence, results in terms of AV evaluation and risk factors may be different for non home-user domains (e.g. industry, government, academia). For example, corporate users may be exposed to different computer threats, or be targeted based on their corporation's characteristics. Fourth, our studied sample is limited to Windows 7 laptops protected by one antivirus product. Hence, our findings do not provide insight into other versions of Windows (e.g. Windows Mobile, Vista, Windows 10, etc.), non-Windows systems such as MacOS and Unix-based OS, other AV products, and other types of devices (e.g. tablet, mobile, desktop).

In addition, our findings may not be representative of other time frames. As security data are known to be dynamic, a similar study conducted at another time-period may lead to different results. This could be particularly true as malware, computer defences, and users evolve over time. Finally, our study was limited to mass market malware attacks. That is, we did not intended to study targeted attacks and zero-day attacks.

## 5.8 Discussion and conclusion

In this article, we presented the results from the first computer security clinical trial of AV software performed with real users in non-laboratory conditions. Similar to clinical trials in medicine, we evaluated the real-life performance of AV software in protecting systems and studied how users interact with the AV, the system and malware attacks as they occurred in the wild. While the studied sample was small compared to medical clinical trials, it is comparable to that of other usability studies and was sufficient to obtain some interesting results with respect to malware attacks risk factors and defence effectiveness.

In terms of AV evaluation, our results show that 38% of users were exposed to a malware attack blocked by the AV, indicating that at least 38% of the users could have got infected had they had no AV installed. In addition, 20% of our users were found to have been infected by some form of computer threats that was not detected by the AV. Interestingly, half of these users did not observe strange behaviour on their laptop even though they were infected. While AV *field efficacy* was estimated at 92%, this performance is below the protection reported by commercial tests for the same product and period. Perhaps this is like vaccine efficacy: since real-life conditions are frequently suboptimal compared with clinical conditions, vaccine protection is often lower than in clinical tests. A similar dynamic may also be taking place with AV product where AV protection is lower with real-life conditions compared to in-lab evaluations Lalonde Lévesque *et al.* (2016a). Finally, the evaluation of the user experience with the AV product revealed variance in results, indicating that one single AV and/or configuration may not accommodate all types of users (Egelman et Peer, 2015).

In terms of risk factors, our results indicate that age, gender, field of expertise, and employment status are not significant correlates of malware attacks. However, we found partial support linking self-reported level of computer expertise to the risk of malware attacks. Users who self-reported high level of computer expertise were found to be more susceptible. Regarding user behaviour, we identified six significant factors; volume of network usage, number and types of web pages visited, default web browser, types of files downloaded from the Internet, and P2P activity. High volume of network usage, and web pages visits were associated with increased risk of malware attacks. We also observed some surprising patterns in web usage, with seemingly innocuous categories of sites such as search engines/portals and computers/Internet being associated with a higher rate of malware attack while more “shady” sites such as those containing pornography content were less so. In addition, using Chrome as default web browser, downloading executables files from the Internet, and engaging in P2P activity were also found to increase the odds of malware attacks. Overall, results suggest that malware attacks may be more a function of frequency and type of online behaviour, rather than based on user characteristics and demographic factors.

Beyond the contribution of these results, this work demonstrates that computer security clinical trials have potential implications for the AV industry. First, it could provide AV testers a viable and complementary approach to tests conducted in controlled environments. Given the realism of the environment and the independence of the malware selection process, tests performed in real-life conditions are less prone to controversy and ethical issues, such as the creation of malware samples. While studies comparing multiple AV or other security products will require more users to get statistically significant results, increasing use of au-

tomation should allow such tests to be performed at relatively modest cost. Second, such studies could be suitable for AV vendors seeking to: i) understand how their products perform in real-world usage, ii) identify which aspects of the product (user interface, detection, remediation, etc.) could be further improved, and iii) identify user groups for which they are more (or less) effective at preventing malware infections. A better understanding of what works best in real-life for specific user groups could help support the design of successful *user-tailored* AV products (Lalonde Lévesque *et al.*, 2016a).

In addition, computer security clinical trials are of potential utility to help understand what user characteristics, demographic factors and behaviour lead to higher risk of malware attacks. This knowledge could be used to improve the content and targeting of user education and training (Oliveira *et al.*, 2017; Lalonde Lévesque *et al.*, 2017), as well as support the development of user risk models for the cyberinsurance industry. To this end, it is important that further research be conducted to assess the multi-level factors of malware attacks. More studies performed in real-life conditions, such as the Security Behavior Observatory (Forget *et al.*, 2014), are needed to validate our findings, and investigate factors that were not included in our study. We hope the work presented here illustrates the merits of future larger scale computer security clinical trials.

**Acknowledgements** This project was funded by Trend Micro and Canada’s Natural Sciences and Engineering Research Council (NSERC), through the Inter-networked Systems Security Network (ISSNet) Strategic Research Network, the Discovery Grant program, and a Canada Research Chair (second author).

## CHAPITRE 6 ARTICLE 3 : AGE AND GENDER AS INDEPENDENT RISK FACTORS FOR MALWARE VICTIMISATION

*Published in the Proceedings of the International British Human Computer Interaction Conference (BHICI) 2017.*

**Authors** Fanny Lalonde Lévesque<sup>1</sup>, José M. Fernandez<sup>1</sup>, Dennis Batchelder<sup>2</sup>

**Institutions** École Polytechnique de Montréal<sup>1</sup>, AppEsteem<sup>2</sup>

**Abstract** This paper presents the results of an empirical study we designed to investigate the independent effect of age and gender as potential risk factors for malware victimisation. Using data collected from Microsoft's Windows Defender on a sample of three million devices running Windows 10, we found that both age and gender are contributing factors for malware victimisation. Men, and young men in particular, were more likely to encounter malware than women, and younger users were more at risk of encountering malware than their older counterparts. However, our findings suggest that the effect of age and gender is not constant across different types of malware. We also discuss potential causes and implications of these age and gender differences in malware victimisation.

**Keywords** Human factor, Computer security, Malware, Field study

### 6.1 Introduction

Human factors (e.g. demographics, characteristics, behaviour) are known to play a significant role in information security. While the literature on user behaviour and cyberattacks is very extensive, there is significantly less work that focus on user demographics. So far many studies have investigated how user demographics relate to cyberattacks; only a few studies have focused on the risk of malware victimisation (Ngo et Paternoster, 2011; Bossler et Holt, 2009; Yen *et al.*, 2014; Lalonde Lévesque *et al.*, 2013). Their findings essentially suggest that age and gender could be contributing factors in the success (or failure) of malware infections.

On the one hand, cybercriminals are increasingly employing varied monetization schemes that target specific regions of the world and categories of users, for example with targeted banking fraud and ransomware attacks. It is conceivable that cybercriminals may be targeting particular groups to maximize success and revenues, in a similar fashion as Internet publicity

campaigns are now targeting specific groups using profiling information based on computer usage behaviour. On the other hand, the psychological traits and level of awareness of users can affect their decision making in the context of computer usage, hence affecting both the likelihood of exposure and the effectiveness of the infection mechanisms. In the first case, there is sufficient circumstantial evidence from the analysis of malware and cybercrime campaigns to believe that users may be targeted according to age and gender. In the second case, previous research has shown that computer usage behaviour varies significantly with age and gender. For these reasons, it is reasonable to hypothesize that age and gender could be actual contributing factors related to the risk of malware victimisation

A better understanding of gender and age differences in the risk of malware victimisation could enable researchers, practitioners and policy makers to better design gender and age-differentiated interventions in cybersecurity. However, rigorous evidence of gender and age differences in malware victimisation are still relatively scarce. Consequently, there is a need to conduct empirical studies of actual malware victimisation based on large and representative sample of computer users. It is therefore essential to try to empirically confirm that age and gender 1) are indeed risk factors, and that 2) they are involved in the causal pathway leading to malware victimisation.

This paper concentrates on the first question, as a precursor for eventually addressing the second one. In particular, we present a large scale empirical study specifically designed to evaluate age and gender as independent risk factors for malware victimisation. Inspired by the epidemiology approach, we design a field study based on a large sample of millions of Windows 10 devices protected by Microsoft's Windows Defender. We use stratification and regression to investigate the effect of age and gender as risk factors of malware victimisation. Our results contribute to existing literature by shedding light on age groups and gender differences in malware victimisation and how their effect vary depending on the type of malware (e.g. ransomware, adware, infostealer).

The remainder of the paper is organised as follows. In Section 6.2 we review previous work on age and gender differences in malware victimisation. Section 6.3 describes the study in terms of design, data collection and analysis. In Section 6.4 we present our results. We discuss our observations in Section 6.5 and limitations of our study in Section 6.6. We conclude and discuss potential implications of our findings in Section 6.7.

## 6.2 Previous studies

There have not been, to the best of our knowledge, other empirical studies specifically designed to evaluate age and gender differences in the risk of malware victimisation. In this section, we present a review of past work that studied how age and gender correlate with malware victimisation; though it was not their primary interest. We also highlight a few studies that investigated the effect of users' demographics on the risk of other types of computer threats (e.g. phishing, spam, identity theft).

### 6.2.1 Demographics and malware victimisation

Some researchers have investigated the effect of users' demographics on malware victimisation by adopting subjective research methods. Mostly based on surveys, interviews, and observations, these methods seek to understand why and how users interact with computer systems. For example, Ngo et Paternoster (2011) applied the general theory of crime and lifestyle/routine activities framework to assess the effects of individual and situational factors on seven types of cybercrime victimization, including computer virus infection. They conducted a self-assessment survey using a sample of 295 college students and correlated users' demographics and characteristics (gender, age, race, marital status) with self-reported cybercrime victimization. The authors deduced that the effect of gender was not significant, while age was identified as a significant predictor for self-reported computer virus infection, with older respondents being less likely to get infected. In another study, Bossler et Holt (2009) applied a routine activities framework to explore the causes and correlates of self-reported data loss from malware infection. The authors administered a survey on a sample of 788 college students and investigated, among others, the effect of gender, age, race, and employment status. They found that being a female increases the odds of malware victimization by 1.827 times compared to male. However, age was not identified as a significant predictor of self-reported malware victimization.

Other studies investigated the effect of age and gender as potential risk factors of malware victimisation based on objective research methods. In comparison with the studies cited above, they are based on real-life data, and not on self-reported malware victimisation and users' behaviour. Lalonde Lévesque *et al.* (Lalonde Lévesque *et al.*, 2013; Lalonde Lévesque *et al.*, 2014) did a 4-month field study of 50 users based on the clinical trial approach used in medicine to assess the impact of human and technological factors on the risk of malware exposure. The authors found no significant differences based on gender or age. Also inspired by the epidemiology approach, Yen *et al.* (2014) conducted a study of malware encounters



in a large, multi-national enterprise. They coupled malware encounters with web activities and demographic information, and found that males were more likely to encounter malware than females.

Although some studies suggest that age (Ngo et Paternoster, 2011) and gender (Bossler et Holt, 2009; Yen *et al.*, 2014) could be significant correlates of malware victimisation, prior work has yielded mixed results in terms of identifying the direction of the aforementioned correlations. For example, Bossler et Holt (2009) found that females are more at risk of malware victimisation, while Yen *et al.* (2014) found that males are at higher risk. Moreover, all research previously cited performed a global analysis of malware, the exception being the work of Ngo et Paternoster (2011) that limited their study to one type of malware (virus). Our research goes beyond as we also evaluate how the direction and magnitude of age and gender vary between different types of malware. Finally, most of these studies offer surprisingly little or no discussion of how the results should be interpreted in terms of causality. In contrast, we also discuss potential underlying causes of how age and gender may affect the risk of malware victimisation—that is whether they have a direct or indirect effect or whether they are confounded by other factors that were not included in our study.

### 6.2.2 Demographics and other computer threats

There is also a number of research that studied the effect of users' demographics on other types of computer threats. For instance, several studies investigated the impact of demographic factors on phishing susceptibility (Sheng *et al.*, 2010; Jagatic *et al.*, 2007; Kumaraguru *et al.*, 2009; Oliveira *et al.*, 2017). Another set of related efforts attempted to examine how demographic factors relate to spam susceptibility (Grimes *et al.*, 2007) or to Internet theft victimization (Reyns, 2013). The overwhelming evidence, however, from all these studies suggests that age and gender are significant correlates for computer threats victimization.

## 6.3 Study design and methods

Since research on demographic factors associated with malware victimisation is relatively sparse, we will derive our hypotheses from past research on demographics and risk in other domains (e.g. finances, career, sports, health). In other words, we are making the assumption that prior studies on age and gender differences in specific domains can perhaps be extrapolated to the risk of malware victimisation. Hence, by extension, we can hypothesize that (H1) gender and (H2) age are independent risk factors for malware victimisation.

### 6.3.1 Case-control study design

We designed a case-control study to test if (H1) gender and (H2) age are independent risk factors of malware victimisation. Commonly used within epidemiology, a case-control study is a type of comparative study where a group of individuals who have a disease (cases) is compared to a group of individuals who do not have the disease (controls). This kind of study is often used to determine whether there is an association between an exposure to a risk (or protective) factor and a disease. In contrast to experimental studies, case-control studies are observational; they do not attempt to alter the course of the disease. Moreover, they are usually, but not exclusively, retrospective by design. They look backwards to learn which individuals in each group (cases and controls) were exposed to the risk (or protective) factor. In other words, once the cases have been identified, the controls are selected from the same population independently of their exposure status.

The frequency of the exposure between the two groups is then compared based on their respective odds of exposure to the potential risk (or protective) factor. From there, the ratio of these odds, the odds ratio (OR), is computed. The confidence interval (CI) in which the true value of the OR is likely to be has to be taken into account when interpreting the OR. An OR larger than 1 indicates that the exposure is a risk factor; the odds of being exposed to the risk factor is higher for the cases than for the controls. To the opposite, an OR smaller than 1 means that the exposure is a protective factor. However, if the OR is equals to 1, or if 1 is included in the CI, nothing can be said on the association between the exposure and the risk of developing the disease.

### 6.3.2 Target population

In order to conduct a case-control study as previously described, we must first select a population on which we will base our study. As our focus is the effect of gender and age as potential independent risk factors (exposure) for malware victimisation (disease), we must also consider any other variables that may affect the risk of malware victimisation. To limit the effect of such extraneous factors that are not of primary interest, we decided to limit our population to one operating system (OS) and one antimalware product. More specifically, our *target population* was limited to Windows 10 devices protected by Microsoft's Windows Defender—an antimalware engine included with Windows that helps detect and mitigate malware on computers. We also added the geographical region where the device is located to control for any potential geographical or cultural effects.

### 6.3.3 Data collection

As our target population was protected by an antimalware product, malware victimisation was computed based on malware encounters reported on devices protected by Microsoft's Windows Defender. As such, we included both known malware attempting to be installed, and malware already installed on the device. Data on malware encounters was collected from October to November 2015 by Microsoft's Windows Defender. Encounters were recorded on all Windows 10 devices that consistently reported with up-to-date antimalware signatures for the entire study; representing a target population of 30+ million devices. All types of devices were included (for example, desktop PCs, notebooks, and tablets) except mobile devices.

Information on those devices was coupled with demographic data from Microsoft Account, a single sign-on web service that allows users to log into various services provided by Microsoft (for example, Outlook, Skype, OneDrive). For each account, associated gender and age group were used. Gender could be male, female or unknown, and age was grouped in six categories (0-17, 18-24, 25-34, 35-49, 50+, unknown). Accounts that had unknown age or gender or more than three devices associated were excluded from the analysis. However, data on malware encounters is collected on the devices and cannot be uniquely associated with a particular user. For example, if the detection happened due to an action by a particular user, e.g. user-triggered scan, the event that initiated the encounter might be attributable to another previous user. To limit this problem, we decided to consider only data from devices that had only one user account. In particular, we excluded devices that had more than one single account associated. In the end, combining the single-account and known-gender/age criteria, we were left with a *sample population* of 3 019 671 million devices. Further, Internet Protocol (IP) geolocation was used to identify the location of those devices. Locations were grouped into the following six categories: North America, Europe, South and Central America, Australia, Asia and Pacific, Africa and Middle East.

### 6.3.4 Ethical and privacy considerations

The telemetry data used in this study was collected by Microsoft in compliance with its security and privacy policies (Microsoft, 2017), as well as international laws and regulations. Data was reported to Microsoft only on devices on which open, or blanket, consent was obtained when installing Windows 10; with the possibility to withdraw, or opt out. For the purpose of this study, only anonymous telemetry data limited to the factors identified in the paper was used.

Table 6.1 Population demographics by factor

Factor	Description	Total population (N=3 019 671)	Case population (N=809 426)	Control population (N=2 210 245)
Gender	Female	25.87%	21.36%	27.52%
	Male	74.13%	78.64%	72.48%
Age	0-17	5.86%	7.27%	5.35%
	18-24	23.03%	29.34%	20.72%
	25-34	25.39%	26.51%	24.98%
	35-49	24.29%	21.17%	25.43%
	50+	21.43%	15.70%	23.53%
Region	Africa & Middle East	3.12%	5.64%	2.20%
	Asia & Pacific	12.19%	16.68%	10.54%
	Australia	2.37%	1.90%	2.54%
	South & Central America	6.82%	11.18%	5.22%
	North America	44.27%	31.25%	49.03%
	Europe	31.24%	33.35%	30.46%

### 6.3.5 Statistical analysis

The risk analysis was determined based on the calculation of the odds ratio (OR), with a confidence interval of 95%. Stratified and multivariate analysis through logistic regression was also performed between the dependent variable (malware victimisation) and the independent variables (age, gender, region). The statistical analysis was conducted using Statistica 12.7.

## 6.4 Results

### 6.4.1 Population

The study lasted two months, and in this period we collected telemetry data from 3 019 671 Windows Defender devices running Windows 10. Table 6.1 presents the basic demographics of each population by factor.

#### Case population

Of the total population, 809 426 devices (26.81%) reported malware encounters during the study. Among all cases, we found a male:female ratio of 3.68:1. Similarly to the total population, the 0-17 age group was the less prevalent with only 58 876 users (7.27%). The distribution between the other groups varied from 15.70% (50+) to 29.34% (18-24). For the region, most of the cases were also either from Europe (33.35%) or North America (31.25%).

## Control population

Of the 2 210 245 devices (73.19%) in the control population, 27.52% were associated with female users and 72.48% with male users. The less frequent age group was 0-17 (5.35%), with other groups ranging from 20.72% (18-24) to 25.43% (35-49). Approximately 80% of controls were either from Europe (30.46%) or North America (49.03%).

### 6.4.2 Malware encounter risk factors

The gender distribution (see Table 6.1) shows that the proportion of male was greater in the case group (78.64%) than in the total population (74.13%); suggesting that being a male may contribute to increase the risk of malware encounter. With respect to the age groups, an increase in the frequency in the case population was seen for the 0-17, 18-24, and 25-34 age groups; indicating that younger users could be more at risk of encountering malware.

### Odds ratio

In order to test the effect of age and gender as risk factors, we computed their respective odds ratio (OR) and confidence interval (CI) at 95%. The effect of gender was investigated with female as a reference level, meaning that male was compared to female. For the effect of age, the age group 50+ was selected as the reference level—all other age groups were tested against this reference. All results in Table 6.2 were statistically significant at  $p$  – value < 0.001 when analysed separately.

Table 6.2 Odd ratios by factor

Factor	Description	OR (95% CI)
Gender	Male	1.40 (1.39-1.41)***
Age	0-17	2.04 (2.02-2.06)***
	18-24	2.12 (2.10-2.14)***
	25-34	1.59 (1.57-1.60)***
	35-49	1.25 (1.24-1.26)***
Region	Africa & Middle East	4.02 (3.97-4.07)***
	Asia & Pacific	2.48 (2.46-2.50)***
	Australia	1.18 (1.15-1.20)***
	South & Central America	3.36 (3.33-3.39)***
	Europe	1.72 (1.71-1.73)***

\*Statistically significant at 0.05 level; \*\*at 0.01 level; \*\*\*at 0.001 level.

Gender was found to be a significant factor associated with malware encounters. More specifically, being a male was identified as a potential risk factor; males were 1.40 times

more likely to encounter malware than females. All age groups were shown to be statistically significant risk factors ( $OR > 1$ ) when compared to the reference level (50+). The groups 0-17 and 18-24 were identified as being the most at risk, followed by the groups 25-34 and 35-49. Overall, results suggest that younger users (0-24) were nearly twice more likely to encounter malware than older users (50+). When analysing if any of the regions were associated with the risk of malware encounter, we found that they were all significant risk factors ( $OR > 1$ ) when compared to the reference region (North America). Africa & Middle East and South & Central America had the highest odds, while Europe and Australia presented the lowest odds. This suggest that all regions are statistically significantly more at risk of malware encounter than North America. Although those results are of inherent interest, the understanding of these geographical variations in malware exposure is out of scope of this paper. Rather, we will focus our analysis and discussion on age and gender variations in malware exposure.

### Stratified and multivariate analysis

To investigate the independent effect of each factor, we used stratification —division of the population in separate groups— to allow the analysis of one factor when controlling for other factors.

Table 6.3 Stratified analysis by studied factors

Risk factor	Stratifying factor	OR (95% CI)
Male gender	Age 0-17	1.53 (1.46-1.60)***
	Age 18-24	1.68 (1.64-1.72)***
	Age 25-34	1.42 (1.39-1.47)***
	Age 35-49	1.17 (1.13-1.21)***
	Age 50+	1.16 (1.11-1.21)***
Age 0-17	Female gender	2.17 (2.06-2.30)***
	Male gender	2.87 (2.79-2.95)***
Age 18-24	Female gender	2.69 (2.69-2.93)***
	Male gender	4.05 (3.97-4.13)***
Age 25-34	Female gender	2.17 (2.07-2.27)***
	Male gender	2.66 (2.61-2.72)***
Age 35-49	Female gender	1.65 (1.57-1.73)***
	Male gender	1.66 (1.62-1.70)***

\*Statistically significant at 0.05 level; \*\*at 0.01 level; \*\*\*at 0.001 level.

Results of the stratified analysis (see Table 6.3 ) support our initial hypotheses that (H1) gender and (H2) age are independent risk factors for malware encounters. Although being a male was identified as an independent risk factor, the magnitude of its effect was smaller for users in the 35-49 and 50+ age groups. The age was also found to be a significant independent

factor after stratification by gender. Interestingly, the impact of age was stronger on male users between 0-34 than on female, which could suggest a potential interaction between the two factors.

Table 6.4 Multiple logistic regression model

Factor	Wald stat.	$p$ – value
Intercept	107 596.70	$< 1.00\text{e-}16$
Gender	4 705.30	$< 1.00\text{e-}16$
Age	26 414.20	$< 1.00\text{e-}16$
Region	88 251.70	$< 1.00\text{e-}16$

A logistic regression model was also developed to study the independent effect of age and gender. This kind of regression was selected as our dependent variable (DV) is binary—it can only take two values. The DV was represented by either 1 or 0, where 1 indicates that the device reported at least one malware encounter over the study. The age was considered as an ordinal discrete independent variable and gender was included as a binary independent variable. Region was also included in the regression as a control variable to account for potential cultural and geographical effects. Similarly to our previous analysis, female gender, age group 50+, and North America were used as the reference levels. We report for each factor the Wald statistic and the  $p$  – value associated. The Wald statistic is used to test the statistical significance of each regression coefficient in the model; the higher the value, the stronger is the effect of the coefficient. The  $p$  – value indicates if the null hypothesis can be rejected, meaning that the coefficient is relevant in the regression model.

Table 6.5 Odds ratios from multiple logistic regression

Factor	Description	OR (95% CI)
Gender	Male	1.24 (1.23-1.25)***
Age	0-17	1.74 (1.72-1.76)***
	18-24	1.78 (1.76-1.79)***
	25-34	1.34 (1.33-1.35)***
	35-49	1.13 (1.12-1.14)***
Region	Africa & Middle East	3.65 (3.60-3.70)***
	Asia & Pacific	2.22 (2.20-2.24)***
	Australia	1.12 (1.10-1.14)***
	South & Central America	3.01 (2.98-3.04)***
	Europe	1.63 (1.61-1.64)***

\*Statistically significant at 0.05 level; \*\*at 0.01 level; \*\*\*at 0.001 level.

Results in Table 6.4 show that all factors are significant at  $p$  – value  $< 1.00\text{e-}16$  in the regression model, which support our initial hypotheses (H1) and (H2). We also computed

from the regression the odds ratio and the 95% CI for each factor. The results in Table 6.5 also confirm that being a male is a risk factor; males were 1.24 times more likely to encounter malware than female. The associations between malware encounters and age groups were also identified as significant risk factors. The odds of malware encounter increase with age until 18-24, after which they decrease; indicating that users in the group 50+ are less likely to encounter malware than the other age groups.

### 6.4.3 Risk factors by malware types

We further wanted to investigate the independent effect of age and gender for different types of malware. Each malware encounter was classified by Microsoft's Windows Defender into a specific malware category. For the purpose of the analysis, malware were grouped in the following 10 categories: adware, virus, cracks, hack, exploit, rogue malware, infostealer, ransomware, bot, and rootkit. See Appendix E for a complete definition of each type of malware. Adware (50.04%) represented half of all the encounters, followed by cracks (16.40%), other (15.75%), and virus (9.40%). All the other categories had proportions smaller than 3%: hack (0.77%), exploit (1.45%), rogue malware (1.85%), infostealer (2.77%), ransomware (0.76%), bot (0.65%), and rootkit (0.16%).

#### Odds ratio

The OR and the 95% CI were computed by studied factors for each type of malware (see Appendix F). Male gender appeared to be a significant risk factor for 8 types of malware: virus, cracks, hack, exploit, infostealer, ransomware, bot, and rootkit. To the opposite, being a male was found to be a weak protective factor for adware encounter (OR=0.98; CI 95%=0.97-0.99); meaning that females were slightly more at risk for this specific type of malware. Moreover, gender was not found to be a significant factor associated with the risk of rogue malware encounter (OR=0.98; CI 95%=0.94-1.02). The same analysis was performed for age by types of malware (see Appendix F). Results show that the effect of all age groups is significant for every types of malware, except for bot and rootkit, where the age groups 0-17 and 35-49 are not statistically significant. Age groups were found to be risk factors — when compared to the reference level (50+) — for 7 types of malware: adware, virus, cracks, hack, exploit, infostealer, bot, and rootkit. To the opposite, all age groups were identified as protective factors for rogue malware and ransomware, meaning that older users (50+) were the most at risk for these specific types of malware. Moreover, results show that the level of risk by age group is function of the type of malware. For example, while users in the 18-24 group are 7.14 times more likely to encounter virus than users in the 50+ group, they are



only 1.36 more likely to encounter adware.

## Multivariate analysis

Similarly to our previous analysis, we conducted a logistic regression in order to study the effect of age and gender as independent risk factors for different types of malware while controlling for potential regional effect.

Results (see Appendix G) show that gender is a significant contributing factor for all types of malware. Interestingly, being a male was found to be a risk factor, except for adware, where it was found to be a weak protective factor ( $OR=0.94$ ;  $CI\ 95\%=0.93-0.95$ ). With respect to age, the effect of all age groups was significant for adware, virus, exploit, and infostealer. However, only one age group was found to be significant for bot and rootkit; suggesting that age may not be an important factor for those types of malware. The odds of infostealer and hack encounters were found to decrease with age. Whereas the odds of virus and cracks encounters exhibited an inverted U-shape trend with age; encounters increase from teenagers (0-17) to young users (18-24), before reducing with age. To the opposite, ransomware and rogue malware encounters were found to increase with age; users in the 50+ age group the most at risk. Hence, hypothesis (H1) is supported for all types of malware and (H2) is only partially supported, as not all age groups were found to be significant.

## 6.5 Discussion

In this section, we give our interpretation of the findings previously reported, focusing on the most interesting results we found. We also compare our results to those reported in prior studies where possible, and highlight instances in which our findings corroborate or refute theirs.

Overall, we found that age and gender are independent risk factors for malware victimisation; males were found to be more at risk of being exposed to malware than females, and younger users at higher risk than older users. As we discuss below, however, the direction and magnitude of the effect of age and gender vary in some surprising ways depending on the type of malware.

### 6.5.1 Gender difference

The risk analysis allowed to identify gender as a significant independent factor related to malware victimisation. Males were found to be 1.24 times more likely to encounter malware

than females. This gender difference was most marked in the population under the age of 25 years, but was also evident among older users. Similarly, Yen *et al.* (2014) found that males were more at risk of encountering malware than females. To the opposite, Bossler et Holt (2009) found that females were more susceptible to malware victimization, as measured by self-reported data loss from malware. However, direct comparison with our results is not possible, as previous work used different study design and target population; Yen *et al.* (2014) studied malware encounters of corporate users within a large enterprise, and Bossler et Holt (2009) based his study on self-reported malware victimisation from college students.

When performing the risk analysis for different types of malware, we also found that being a male was a risk factor, except for adware. For this specific type of malware, being a male was a weak significant protective factor; meaning that females were slightly more susceptible to encounter adware than males. We present in the following text potential underlying causes that could explain such difference across gender and types of malware.

### **Risk attitude**

A first possibility for this gender difference could be that males are more susceptible to malware victimisation than females because of their attitude towards cybersecurity-related risk. This could be plausible as gender differences in risk attitude has been identified across various contexts, such as car driving, financial matters, health, social decisions, sport and leisure, and career (Byrnes *et al.*, 1999; Weber *et al.*, 2002; Dohmen *et al.*, 2011; Harris *et al.*, 2006). Though there is extensive evidence to show that males are more risk seeking than females overall, the direction and magnitude of the gender effect tend to depend of the domain. For example, while male are more likely to exhibit risky behaviors in car driving, researchers found that female report greater propensity than male to engage in risky behaviors when it comes to social decisions Weber *et al.* (2002); Johnson *et al.* (2004). These variations across domains have been attributed, among others, to gender differences in (1) the perceived probability of negative consequences, (2) the perceived severity of a potential negative consequences, and (3) the enjoyment of engaging in risky behaviors (Harris *et al.*, 2006; Emond *et al.*, 2009; Wang *et al.*, 2011; Hogarth *et al.*, 2007). However, the extent to which they are the product of genetic, social, developmental, or experimental factors is still lacking strong consensus in the research literature.

Similarly, one could argue that males are more likely to encounter malware than females because of gender differences in risk perception and enjoyment of risky behaviors in cybersecurity. This could imply that (1) males have lower perceptions of the probabilities and severity of negative consequences from engaging in risky behaviors in cybersecurity, and (2)

they expect higher enjoyment than females from these behaviors. Those gender differences could explain, for example, why males were found to be 1.65 times more likely to encounter cracks –tools often used to engage in software piracy– than females.

### Computer usage

With car driving, we know that both driving behaviors and time spent on the road are significant contributing factors to the risk of car injury. Similarly, a second explanation could be the difference in frequency and type of computer usage behavior between male and female (Hu *et al.*, 2007; Joiner *et al.*, 2012; Goel *et al.*, 2012). This is consistent with previous work that identified gender differences in frequency and patterns of Internet use. For example, Joiner *et al.* conducted a survey of 501 students and found males to be heavier Internet users than females (Joiner *et al.*, 2012). Males were more likely to use the Internet for games and entertainment, to bet online, to visit web sites with adult content, and to download music and videos. On the other hand, females were more likely to use the Internet for communication (e.g. email, telephone), and visit social network sites. In another study, Goel *et al.* examined the Web histories of 250,000 anonymized individuals paired with user-level demographics. They found that females spend considerably more time online on social media sites, and that visits to sports sites are highly predictive of being male (Goel *et al.*, 2012).

Moreover, several research found empirical evidence of associations between the frequency and type of computer usage and the risk of malware exposure. Carlinet *et al.* (2008) performed a case-control study based on network traffic of a large set of real ADSL customers. They found that surfing the web a lot and high usage of streaming applications are risk factors to being infected with malware. In another study, Lalonde Lévesque *et al.* (Lalonde Lévesque *et al.*, 2013; Lalonde Lévesque *et al.*, 2014) found evidence that installing many applications and visiting many web sites may increase the risk of malware encounters. They also identified specific categories of web sites, most of which were legitimate, that were more likely to be associated with increased risk of malware encounters. Similar results were also obtained by Canali *et al.* (2014). The authors developed a risk model of malware encounter based on users' web browsing behavior. They used a large telemetry dataset collected by a major antimalware vendor and identified specific web sites categories, and the total number of web sites visited, as good predictors of the likelihood of encountering malware. This last finding was also supported by the work of Yen *et al.* (2014), which identified a positive correlation between the volume of user activity (as measured by the number of distinct domains visited by a host) and the probability of encountering malware.

Overall, the studies cited above support the existence of a relationship between web browsing behavior and the risk of malware victimisation. This trend is consistent with recent observations and reports by the antivirus (AV) industry. In particular, a recent report by Microsoft (Anthe et Chrzan, 2015) identifies *web browsing* as being the most frequent transmission vector used by malware for the first quarter of 2015 (Anthe et Chrzan, 2015), the period just 6 months ahead of our study. Although results are not limited to Windows 10 users, they provide strong evidence that most users encountered malware because they either visited a malicious or compromised web page, or downloaded a malicious application (voluntarily or not). For example, users can get infected through malvertising —malicious advertising— by clicking on an innocuous-looking banner ads containing malicious code (Sood et Enbody, 2011; Xing *et al.*, 2015). Other attacks, such as drive-by downloads (Mavrommatis et Monrose, 2008; Provos *et al.*, 2007), can download malware without any user intervention required, by either operating malicious web sites or by injecting malicious content into compromised legitimate web sites. Finally, users can also get infected by downloading a piece of software (e.g. free games, media players, screen savers, keygens) that comes bundled with spyware, adware or malware.

In light of this discussion, males could be more at risk of encountering malware than females because (1) they are heavier computer users (e.g. they visit more web sites, they install more applications), and (2) they are more prone to engage in computer behaviors that may, intentionally or not, increase their likelihood of encountering malware. Similarly, females could be more at risk of adware encounter as a result of differences in their computer behavior (e.g. categories of web sites visited, type of applications installed). Although these hypotheses are plausible, additional research should be conducted in order to gain a better understanding of how computer usage behavior affect the risk of malware victimisation, and establish sound causation.

### 6.5.2 Age difference

Results suggest that age is a significant independent risk factor for malware victimisation. Young users (0-24 years), in particular users in the 18-24 age group, were the most likely to encounter malware. To the opposite, older users (50+) were found to be the less susceptible to encounter malware. This supports the findings of Ngo et Paternoster (2011) that suggest that older users are less likely to get infected by malware.

Our risk analysis by types of malware reveals, however, that the direction and magnitude of the age effect is a function of the type of malware. Although increasing age was associated with reduced malware encounters overall, its effect was particularly strong for virus and

infostealer encounters, and relatively small for bot and rootkit encounters. Moreover, while older users (50+) were found to be less at risk of encountering malware overall, they were the more susceptible to encounter rogue malware and ransomware. We present in the following text potential causes for these age differences.

## Risk attitude

Similarly to gender, age differences in malware victimisation could be attributed to variations across age groups in risk attitude towards cybersecurity. In comparison, age differences in risk-taking behaviors have also been identified in multiple risk domains (Rolison *et al.*, 2013; Dohmen *et al.*, 2011, 2005). There is an overwhelming consensus that young age is associated with higher willingness to take risks than older age (Dohmen *et al.*, 2005). However, studies also reveal that age differences in risk-taking may depend on the domain. For example, Rolison *et al.* (2013) found that risk taking in the financial domain reduces steeply with older age, while in the social domain, it increases slightly from young to middle age, before reducing sharply in later life. A number of possible underlying causes, such as (1) changes in life circumstances, (2) motivational factors, and (3) cognitive decline, have been advanced to explain such variations (Rolison *et al.*, 2013; Mather, 2006). While these causes might be relevant for risk tendencies in specific domains (e.g. financial, social, recreational), risk attitude in cybersecurity may differ, and point to different underlying causes. As with gender, we believe that age differences in malware victimisation may be, to some extent, attributed to age changes in risk perceptions and expected enjoyment of engaging in risky behaviors. Specifically, this could imply that (1) younger users have lower perceptions of the probabilities and severity of negative consequences from engaging in risky behaviors in cybersecurity, and that (2) they expect higher enjoyment than older users from engaging in risky behaviors.

Another possibility could be that malware encounters differ across age groups as a result of changes in emotional processing. This is consistent with previous research in psychology, sociology and economics, that identified emotion to be a major determinant of risk perception and risk taking that changes with age (Figner *et al.*, 2009). While emotions are found to act as an *advisor* for risk taking in situations of low level of emotional intensity, they seem to inhibit cognitive processes in situations of high level of emotional intensity (Bieberstein, 2013). Emotional differences could therefore explain why older users (50+) are more likely to encounter rogue malware and ransomware than younger users. As those categories of malware are known to use deceptive fear to trick users into downloading a malicious software (a trial version of a bogus security software or a fake software update), older users could be more likely to act by emotions rather than by cognitive processes when exposed to such

trickery. Hence, older users would be more susceptible to rogue malware and ransomware because of emotional differences when faced with persuasive messages that attempt to scare them.

### Computer usage

Another likely reason could be age differences in frequency and type of computer usage. This is supported by prior studies that identified differences in volume and type of computer activities across age. By analysing the web histories of 250,000 individuals, Goel *et al.* (2012) found that younger users spend much more time online relative to their older counterparts. Their results also reveal that older users spend a smaller fraction of their online time on social media web sites. In another study, Teo (2001) conducted a web-based survey of 1,370 respondents to examine how demographics variables and motivation variables correlate with Internet usage activities (messaging, browsing, downloading, and purchasing). Their results show that younger users engage in messaging and downloading activities to a greater extent than older users.

Taken together with previous findings of the relationship between computer usage and risk of malware victimisation (as presented in Section 6.5.1), we can hypothesize that younger users could be more likely to encounter malware because (1) they are heavier computer users, and (2) they engage in computer activities that could contribute to increase (intentionally or not) their risk of malware victimisation. Furthermore, older users could be more likely to encounter rogue malware and ransomware because of their computer activities. This is possible, as rogue malware and ransomware are known to target specific countries, OSes, programs, companies, or web site categories. Similarly, older users could engage in computer activities (e.g. visiting specific categories of web sites, installing/using specific types of applications) that would increase their likelihood of encountering such attacks. However, the extent to which older users are more exposed as a result of their computer activities, or because they are seen as attractive targets (lack of Internet savvy, potential access to life savings, and impaired decision making due to ageing) remains unknown.

### 6.5.3 Summary of findings

We presented in this paper a number of interesting findings related to gender and age differences in the risk of malware encounters. The key findings of our study can be summarized as follow:

- Age and gender are significant independent factors of malware encounter.

- Male, and young male in particular, are more likely to encounter malware than female.
- Female are slightly more at risk of encountering adware than male.
- The gender difference is most marked in the population under the age of 25 years, but is also evident among older users.
- Increasing age leads to decreasing risk of malware encounter; younger users (0-24) are more at risk of encountering malware than older users (50+).
- Older users (50+) are the most susceptible to encounter rogue malware and ransomware.

## 6.6 Study limitations

Although case-control studies allow determination of whether an exposure is associated with an outcome, their results can be highly sensitive to bias, confounding variables, and chance circumstances. Hence, our study and its conclusions are subject to a number of limitations and potential bias that may affect its internal and external validity. Internal validity refers to the strength of the inferences from the study, that is the extent to which no other variables except the one we studied caused the results. While external validity refers to the ability to generalize the results to a more *universal population*.

First, malware encounters are limited to the malware families detected by Microsoft's Windows Defender. While these malware may represent some of the most significant malware families on Windows, they do not cover targeted attacks and zero-day attacks. Moreover, the encounters reported depend on the efficacy of Windows Defender, which may lead to an underestimation of malware encounters. Nevertheless, given the significance of the malware families covered by Windows Defender, these encounters are also of inherent interest, whether or not they are representative of all computer threats on Windows 10.

Second, the sample population is limited to devices that have known age/gender and a single-account associated. Hence, the exclusion of devices with multiple accounts, or with missing demographic information may have introduced a sampling bias. In order to estimate this potential bias, we compared our sample population (3+ million) against our target population (30+ million). We found that both populations were similar in terms of geographical distribution and malware encounters. However, this does not imply that the two populations are similar in terms of other factors, such as demographics or risk attitude. Moreover, our sample population may not be representative of the target population for other time frames. As security data are known to be dynamic, a sample population drawn from the same target population at another time-period may be different. This could be particularly true as our study was conducted few months (Oct.-Nov. 2015) after the official release of Windows

10 (July 2015); meaning the target population may evolve over time as more users adopt Windows 10.

Another limitation of our study is its susceptibility to confounding. Although the region factor was included in our analysis to account for potential geographical or cultural effect, and multivariate analysis was used, we cannot guarantee that our results were not affected by other unknown extraneous variables that may confound the results. For example, it may be possible that in some cases several human users shared the same user account on the same single-account device, which may have introduced a bias that we were not able to control nor measure. It would be interesting in future work to consider additional extraneous variables, such as education or social status.

Finally, a significant limitation to our external validity derives from our target population –Windows 10 devices protected by Microsoft’s Windows Defender. As our analysis was limited to Windows 10 devices, it does not provide insight into other versions of Windows (e.g. Windows Mobile, Vista, XP, etc.), and it does not give insight into the encounter rates on non-Windows systems such as MacOS and Unix-based OS. Furthermore, the analysis was limited to Windows 10 devices running Windows Defender. Thus, it does not cover users protected by other antimalware products. However, given that Defender was running on more than 40% of all Windows 10 devices during the period covered by our study, we believe our findings are important on their own, whether or not they are representative of patterns in devices protected by other antimalware products. Though we agree that a study including multiple antimalware products is interesting and would provide additional insights, such analysis was outside the scope of this study.

## 6.7 Conclusion

We presented the results of a large scale empirical study specifically designed to evaluate gender and age as potential independent risk factors of malware victimisation. While our work corroborates some findings in earlier research, our results support our initial hypothesis, that both (H1) gender and (H2) age are contributing independent factors correlated to the risk of malware victimisation. Those results were also robust after stratification and multivariate analysis. Male and younger users were found to be more at risk of malware encounter overall, though the direction and magnitude of the gender and age differences varied depending on the type of malware. Interestingly, certain types of malware were associated with nontrivial age differences (e.g. ransomware and rogue malware), whereas others were associated with gender differences that shifted from risk factor to protective factor (e.g. adware).



It is clear from the evidence that differences between the age groups and gender exist in the context of malware victimisation. The remaining question concerns the origins of these associations, i.e. their causality. We have discussed potential underlying causes that could explain why age and gender are risk factors. In particular, we hypothesize that differences in attitude towards risk taking and differences in computer and Internet usage, which have been reported to change with age and gender, could explain the differences in malware victimisation. Verifying these causal hypotheses is essential for the design of successful targeted, age and gender differentiated interventions aimed at preventing or reducing the risk of malware victimisation.

In particular, this study and its findings may help support the development of user-differentiated human-computer systems. As systems designed to suit the *average user* may not accommodate all user groups (Egelman et Peer, 2015), security systems could be tailored to users' risk of victimisation. For instance, one recent study provided preliminary evidence that antivirus effectiveness differs significantly across demographic factors; antivirus had lower performance for female users and the 0-17 age group (Lalonde Lévesque *et al.*, 2016b). Demographic factors could then be used to infer the risk of malware victimisation, and personalize systems (default security settings, human-computer interfaces, etc.) in order to maximize protection for all user groups.

In addition, this could have potential implications for the cyberinsurance industry as well. For example, in the car insurance industry personal characteristics such as age, gender, and marital status are often used as proxies of driver behavior (accelerating, braking, etc.) and driving characteristics (where, when, etc.). Although finer-grained data on driving can be collected through vehicle telematics, the use of such devices is not always available for drivers and insurers. Besides, the collection of such data brings up a number of privacy concerns and other ethical issues, especially concerning computer and Internet behaviour and usage, which is potentially much more privacy-invasive than driving data. Thus, we believe that it could be useful to develop predictive user risk models that use coarse non-invasive information, in order to address these privacy concerns, while supporting the risk-selection needs of the cyber insurance industry.

Furthermore, more studies are needed based on alternate observational data sources, other time frames and different analysis methods in order to confirm that our findings are robust across different populations.

Finally, we believe it is important to try to identify and validate the potential causality of other risk factors that may be associated with malware victimisation, such as other personal traits, and socio-economical and cultural factors. Determining *who* is more susceptible to

malware victimisation and *why* is paramount to improve security for *all* users.

**Acknowledgements** The authors would like to thank the Microsoft Malware Protection Center (MMPC) for supporting this work and granting us access to the Microsoft's Windows Defender telemetry data.

## CHAPITRE 7    ARTICLE 4 : MEASURING THE HEALTH OF ANTIVIRUS ECOSYSTEMS

*Published in the Proceedings of the 10th International Conference on Malicious and Unwanted Software (MALWARE) 2015.*

**Authors**    Fanny Lalonde Lévesque<sup>1</sup>, Anil Somayaji<sup>2</sup>, Dennis Batchelder<sup>3</sup>, José M. Fernandez<sup>1</sup>

**Institutions**    École Polytechnique de Montréal<sup>1</sup>, Carleton University<sup>2</sup>, AppEsteem<sup>3</sup>

**Abstract**    The number and variety of computer threats has fueled a digital arms race, resulting in a complex software ecosystem around malware and antivirus (anti-malware) products. While there has been significant past work in benchmarking antivirus (AV) products against each other, how healthy is the overall AV software ecosystem?

Using data collected from Microsoft Windows Malicious Software Removal Tool (MSRT) running on more than one billion machines, we develop ecosystem health measures based upon infection rates, product diversity, market dominance, and activity status. Our study shows that while a diverse group of products is used and the vast majority of them are running properly, there is also significant churn in product usage which may indicate dissatisfaction with current products. While further work is needed to better understand these patterns, this study shows the potential power of an ecosystem health-based approach to studying AV performance in practice.

### 7.1 Introduction

The multitude of security products available on the market has evolved into a complex ecosystem that interacts with the malware landscape. Given the increase in complexity and diversity of both malware threats and antivirus (AV) products, the evaluation of the latter is essential in helping the industry develop better products that match the evolving nature of malware and meet users' expectation.

While typical evaluation methods are mostly focused on single-product or comparative tests, AV products are always evaluated on an individual basis and not as a whole. Measuring the overall performance of AV products can provide a better understanding of their global condition and help identify issues that could not be studied using current AV testing methods.

Moreover, such evaluation can allow the investigation of the aggregated effect, if any, of AV products beyond their individual contribution.

One approach is to consider AV products as software ecosystems, that is as a collection of software solutions that are developed and co-evolve in the same environment. The concept of *ecosystem health*, which refers to the global condition of an ecosystem, provides a powerful theoretical and practical framework for monitoring system activity, identifying and predicting areas for improvement, and evaluating changes in ecosystems (Bertollo, 1998). Applied to AV products, the “health” of AV ecosystems can be measured as its overall performance, that is how well it is protecting users against specific, prevalent malware. Developing relevant indicators such as the number and relative usage of different AV products, or how well maintained those installations are, could allow to track the status of the AV ecosystem and assess its overall condition and quality.

In this paper we report on the first study that aimed to define and measure the health of AV ecosystems by developing scalable indicators in terms of activity, diversity and stability. Using four months of sampled telemetry data from Microsoft Software Removal Tool (MSRT) on millions of computers, we analyse AV product status (whether they are running and have up to date signatures), AV vendor diversity, and AV stability in terms of protection status and security vendor. We examine some initial testing to investigate how those indicators relate to MSRT infection rates (of malware missed by the installed AV), discuss opportunities for future testing and successfully identify areas that could be improved within the AV ecosystem.

The remainder of the paper is organized as follows. Section 7.2 presents the literature and related work in natural and software ecosystems health. In Section 7.3 we describe our indicators of AV ecosystems and discuss their associations with users’ protection in Section 7.4. We discuss methodological limitations and future work in Section 7.5 and conclude in Section 7.6.

## 7.2 Background

First introduced in the field of natural ecosystems, Costanza *et al.* (1992) defines a healthy ecosystem as being “stable and sustainable”. While there is no universally accepted definition and indicators, ecosystem health could be defined as a combined measure of system vigor (productivity), organization (including diversity and interactions) and resilience. Vigor or productivity refers to the capacity of the system to sustain its activity. Organization refers to number and diversity of interactions between components of the system. Resilience refers to the ability of the system to maintain its structure and activity in the presence of stress. That

is, a healthy natural ecosystem is one that can develop an efficient diversity of components and exchange pathways (organization) while maintaining its activity (vigor) over time in the face of stress (resilience).

Beyond natural ecosystems, the notion of ecosystem health has also inspired the field of software ecosystem (SECO), where health refers to how well the ecosystem is functioning, that is its ability to endure and remain variable and productive over time (Manikas et Hansen, 2013). For example, Wynn Jr *et al.* (2007) applied the concept of natural ecosystem health to develop a framework in terms of vigor, resilience and organization to gauge the health and sustainability of open source projects.

SECO health has also been applied with a business ecosystem (BECO) approach, as a means of expanding development, better positioning in the market, or increasing revenues. In the BECOs health literature, the concept of health is defined as the ability of a BECO to provide "durably growing opportunities for its members and those who depend on it" (Iansiti et Levien, 2004a). The notions of vigor, organization and resilience are adopted and changed to productivity, niche creation and robustness (Iansiti et Levien, 2004a,b; Iansiti et Richards, 2006).

While ecosystem health has been applied to various SECOs, such as open source software (Jansen, 2014; Van Lingen *et al.*, 2013) or hardware-dependent software (Wnuk *et al.*, 2014), it has not been used significantly in the area of AV software. Many have talked about ecosystems and ecosystem-related concepts in computer security, particularly with regards to monocultures (Geer *et al.*, 2003) and mechanisms for automated software diversity (Forrest *et al.*, 1997). When assessing the performance of AV systems, however, the main focus has traditionally been on single-product or comparative tests of AV systems' ability to detect malware and ignore benign software. Whether performed in a controlled lab environment or through field studies (Somayaji *et al.*, 2009; Lalonde Lévesque *et al.*, 2013), current evaluation methods are limited to the individual performance of security products.

Security vendors have also used software telemetry data for quality assurance. However, those analysis are not intended to study the overall performance of AV systems, but rather focus on one single vendor. Closer to our research is the work done by Blackbird et Pfeifer (2013a), where they used MSRT data to evaluate the global impact of anti-malware protection state on infection rates of protected users. To the best of the authors' knowledge, there has been no previous work published in the literature based on such telemetry data to assess the overall health of AV ecosystems. The key contributions of this work are therefore the proposing of measures of AV ecosystem health and assessing that health using large-scale security software telemetry data.

### 7.3 Antivirus ecosystem indicators

The evaluation of AV ecosystems requires the development of relevant, scalable, easy to measure and understand indicators. In this work we propose to characterize AV ecosystems in terms of activity, diversity, and stability. Using our indicators, we conduct a longitudinal analysis to track the status of the global AV ecosystem over a 4-month period.

The data was collected by MSRT, a malware cleaner utility that scans computers for infections of specific, prevalent malicious software and helps remove these infections. MSRT is delivered and runs every month on more than one billion machines through Windows Update as well as being available as a separate download from Microsoft. It is worth mentioning that MSRT only runs on Windows and that it, by design, only detects a subset of the malware families covered by Windows Defender and other Microsoft anti-malware products. Upon its execution, MSRT also calls the Windows Security Center (WSC) API to collect information about the protection state of computers, such as the AV actively protecting the machine and its signature status. The data used in this analysis was collected from June 2014 to September 2014 on computers running Windows XP, Vista, 7, 8 and 8.1. As not all Windows users send their data to MSRT, we randomly selected one on every ten unique computers in order to limit potential effect of self-selection bias, reducing our sample size from one billion to 100+ million computers. Moreover, we restricted our analysis to the approximately 90% of computers that had an AV product installed, giving us a sample population of 90+ million hosts.

#### 7.3.1 Activity

We define the activity of the AV ecosystem as the percentage of users with at least one AV product actively running with up to date signatures. Figure 7.1 illustrates the evolution of the percentage of users having an up to date AV installed. Over the studied period, the activity of the AV ecosystem ranged from 87.50% to 88.60%.

Table 7.1 AV status over the 4 months

	Jun.	Jul.	Aug.	Sept.
Enabled	88.46%	87.99%	87.80%	88.60%
Out of date	3.74%	3.91%	3.96%	3.74%
Expired	1.87%	2.23%	2.22%	1.98%
Snoozed	0.45%	0.43%	0.33%	0.30%
Off	5.48%	5.44%	5.70%	5.38%

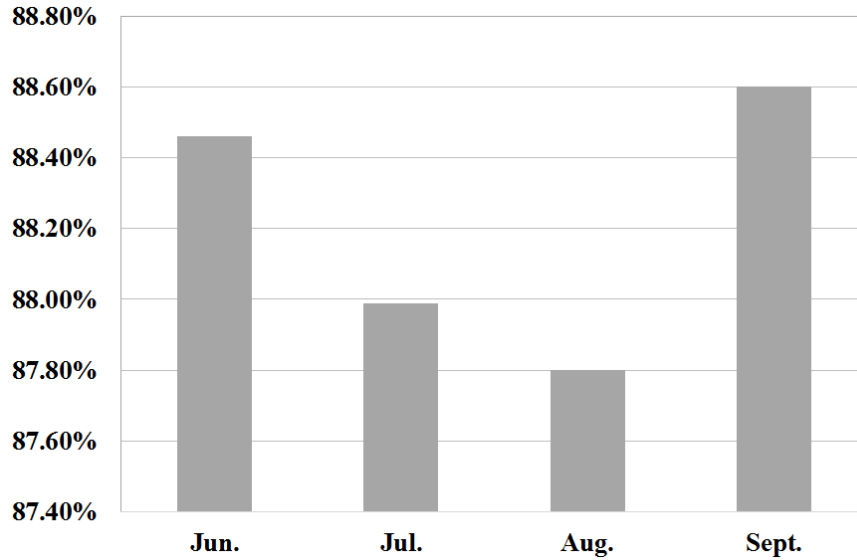


Figure 7.1 AV activity over the 4 months

More details on the status of AV products during the 4 months are presented in Table 7.1. The status enabled refers to an actively running AV product using the latest signature files available while out of date refers to an actively running AV product using out of date signatures. Expired refers to an actively running expired AV product. Snoozed means that the AV product is active but is not performing real-time monitoring, typically because the product is upgrading itself, and off means that the AV product is not running, as it has been turned off. Table 7.1 shows that within the users that have an AV product installed, between 11.40% and 12.21% are not protected with up to date signatures, despite having an AV installed.

### 7.3.2 Diversity

The diversity of the AV ecosystem was characterized based on its richness, degree of concentration, and dominance.

In natural ecosystems, the richness (S) refers to the total number of different species. Applied to the global AV ecosystem, the richness can be defined as the total number of AV vendors within the ecosystem. Table 7.2 shows the evolution of the different indicators in terms of diversity. Over the 4 months, the richness did not vary from 107, indicating that the number of AV vendors was constant.

The degree of concentration (D), also known as Simpson's diversity Index Simpson (1949) or Gini-Simpson Index, is a measure of the degree of concentration of individuals classified

into types. It can be interpreted as the probability that two organisms belong to different species. The value  $S$  represents the richness –the total number of different AV vendors– and  $p_i$  represents the fraction of AV products that belong to the  $i$ th AV vendor. A value of 0 indicates no diversity and 1.0 indicates high diversity.

$$D = 1 - \sum_{i=1}^S p_i^2$$

Based on the results in Table 7.2, we can tell that the AV ecosystem is highly diversified, as its degree of concentration varies around 0.92.

Dominance of the AV ecosystem was measured using the Berger-Parker Index (BP) (Berger et Parker, 1970). This index estimates dominance using the prevalence of the most abundant type, which refers to the AV vendor with the highest market share. Results in Table 7.2 show that the dominance of the AV ecosystem varies between 0.136 and 0.143. In economics, that would indicate that the AV market has a low concentration ( $<0.5$ ), ranging from perfect competition to an oligopoly.

Table 7.2 AV diversity over the 4 months

	Jun.	Jul.	Aug.	Sept.
Richness	107	107	107	107
Concentration	0.906	0.906	0.906	0.928
Dominance	0.176	0.179	0.136	0.181

### 7.3.3 Stability

Stability of the AV ecosystem was evaluated in terms of changes in AV status and AV vendor. Table 7.3 shows for each month the percentage of users by status that have a different AV status compared to the previous month. For example, the value in the first line (enabled) under the column Jul. means that 3.19% of the users that had an enabled AV product for June had a different AV status for July. The next value in the same line means that 3.57% of the users that had an enabled AV product for July had a different status for August. Interestingly, the rate of changes are not equivalent between the different AV status. Users with snoozed AV products, followed by out of date AV products, are the status with the highest rate of changes. The monthly rate of AV state changes varies between 10.84% and



11.91% (Table 7.3) and overall, 40.47% of the users changed their AV status over the 4 months.

In order to better understand the nature of those changes, we analysed for each month the percentage of users that switched from one AV status to an other. We found that more than 75% of the variations for enabled AV products went to either an out of date or off AV status. For all the other statuses (out of date, expired, snoozed, off), 75% of all changes in AV status went to an enabled AV status.

Table 7.3 AV state changes over the 4 months

	Jul.	Aug.	Sept.
Enabled	3.19%	3.57%	3.16%
Out of date	32.25%	36.05%	37.90%
Expired	18.59%	15.53%	23.14%
Snoozed	49.62%	53.81%	45.76%
Off	19.85%	20.51%	25.92%

We adopted a similar approach to evaluate the stability of the AV ecosystem in terms of changes in AV vendors. Overall, 33.57% of the users switched to a different AV vendor over the study (Figure 7.1).

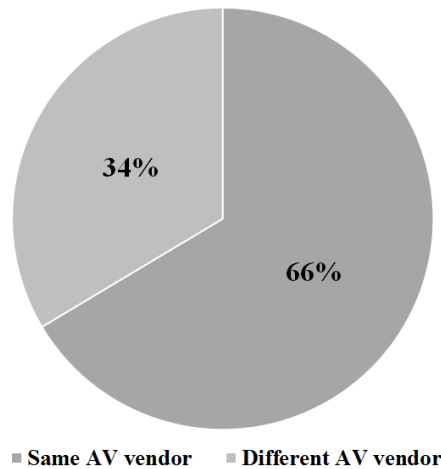


Figure 7.2 Overall AV vendor changes

We also investigated potential relationships between changes in AV status and AV vendors. As presented in Table 7.4, we can see that stability in AV vendors is associated with higher stability in AV status. To the opposite, 87.10% of the users that changed their AV vendor also experienced changes in the status of their AV product.

Table 7.4 Overall AV stability

	Stable Status	Different Status
Stable AV vendor	83.09%	16.91%
Different AV vendor	12.90%	87.10%

#### 7.4 Country level analysis and evaluation

We define the health of the AV ecosystem as the measure of its aggregated performance, that is how well it is protecting users. As illustrated in Figure 7.3, we can see that the overall infection rates of users that had an AV product installed depend of the protection status of the latter. Not surprisingly, users with an enabled AV products have the lowest infection rates.

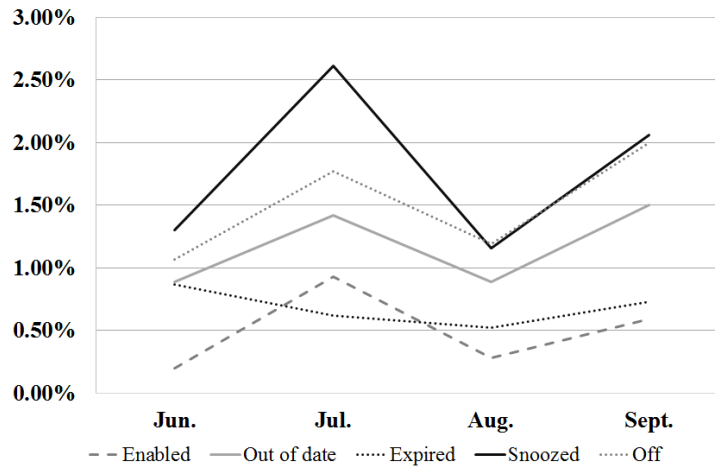


Figure 7.3 Infection rates over the 4 months

To evaluate how our indicators relate to users' protection in terms of infection rates, we conducted an empirical study of AV ecosystems defined by geographical unit, as classified by MSRT. While our data set is large overall, for some countries our sampled population is too small to allow for proper analysis. To determine the minimum representative sample size for each country, we performed a power analysis. We used a two-tailed one proportion Chi-Square test with a desired power of 90% and a level of significance of 1%. The minimum sample size computed was 37 149, which can be rounded to 38 000. We then excluded all countries that had less than 38 000 reports over the 4 months, reducing our sample from 187 to 126 countries.

Our sample of MSRT data contained many users that only report data sporadically. Because

we are interested here in health changes over time, we limited our analysis to the portion of users that reported for all four months of the study period. This exclusion removed roughly two-thirds of the sample, leaving us with approximately 32+ million users.

Infection rates by country were used as the dependent variable in order to estimate the aggregated performance of AV ecosystems. The independent variables were selected to capture the activity, diversity and stability of AV ecosystems. The activity was represented by the proportion of users that had an actively running AV product with up to date signatures for the entire period of the study. The diversity was evaluated based on the 4-month averaged richness, degree of concentration, and dominance. The stability was computed by the proportion of users that experienced changes in AV status and the proportion of users that changed AV vendor during the studied period. Descriptive statistics for each factor are presented in Table 7.5. The mean allows to measure the central tendency of the data and the standard deviation measures how concentrated the data are around the mean; the more concentrated, the smaller the standard deviation. From Table 7.5, we can see that richness varies widely across the different ecosystems. To the opposite, %Protected and concentration present small dispersion.

Table 7.5 Descriptive statistics

Dimension	Indicator	Mean	Std. Dev.
Activity	%Protected	0.849	0.049
Diversity	Richness	50.14	14.01
	Concentration	0.827	0.040
	Dominance	0.296	0.067
Stability	%Unstable AV status	0.482	0.091
	%Unstable AV vendor	0.399	0.085

In order to measure potential dependence between the dependent variable and our indicators, we first computed the Pearson correlation coefficients between the indicators and the infection rates (see Table 7.6). The value  $r$  represents the correlation coefficient. A value of 1 implies that there is a linear relationship between the two factors; as one increases, the other also increases. To the opposite, a value of -1 means that when one value increases, the other decreases. And a value of 0 indicates that there is no linear correlation between the variables. The p-value was also computed to measure the significance of the results. A low p-value (such as 0.01) means that there is a 1 in 100 chance that we would have obtained the same results if the variables were not correlated. For the purpose of our analysis, we considered a correlation to be significant if the p-value was lower than 0.01. Results show that activity, diversity in terms of richness and stability are significantly correlated with the infection rate (p-value<0.001). Activity and richness are found to be negatively associated with the

infection rate, meaning that high values are associated with low infection rates. To the opposite, higher changes in AV status or AV vendors are associated with higher infection rates.

Table 7.6 Pearson correlation coefficients between infection rate and indicators (N=126 countries)

Indicator	<i>r</i>	p-value
%Protected	-0.59	2.14e-13*
Richness	-0.42	6.99e-07*
Concentration	-0.08	3.70e-01
Dominance	0.15	8.51e-02
%Unstable AV status	0.71	7.03e-21*
%Unstable AV vendor	0.67	8.16e-18*

Although the Pearson correlation coefficient provides insight on the dependence between the infection rates and the indicators, its primary drawback is that it is very difficult to draw conclusions about the effect of one single factor on the dependent variable, as factors often interact together. We therefore conducted a multiple regression to examine the relative importance of each indicator. Multiple regression was selected as it allows to estimate the effect of the factors while controlling for the many factors that simultaneously affect the dependent variable. Because we are interested to assess the unique effect of each indicator, we looked for multicollinearity as it can reduce the effective amount of information available to evaluate the effect of the indicators. The presence of multicollinearity was investigated by computing the Pearson correlation coefficient matrix. Results in Table 7.7 show the presence of a very strong correlation ( $r > 0.90$ ) between the two indicators related to stability. We therefore excluded the indicator %Unstable AV status from our analysis and only kept %Unstable AV vendor to estimate the stability of AV ecosystems.

Table 7.7 Pearson correlation coefficient matrix between indicators (N=126 countries)

	%P	R	C	D	%UAVS	%UAVV
%Protected (%P)	1.00	0.38***	-0.08	0.10	-0.69***	-0.60***
Richness (R)	-	1.00	0.05	-0.05	-0.05***	-0.49***
Concentration (C)	-	-	1.00	-0.90***	-0.05	-0.05
Dominance (D)	-	-	-	1.00	0.09	0.07
%Unstable AV status (%UAVS)	-	-	-	-	1.00	0.98***
%Unstable AV vendor (%UAVV)	-	-	-	-	-	1.00

\*Statistically significant at 0.05 level; \*\*at 0.01 level; \*\*\*at 0.001 level.

Table 7.8 presents the results from the multiple general linear regression. For each factor, the standardized regression coefficient  $\beta$  and its associated standard error (Std. Error) were

computed. The p-value, which is interpreted as an indicator of the significance of the results, was also computed: a low p-value indicates that the null hypothesis can be rejected with high confidence, and that the variable is relevant in the regression model. In order to limit potential effect of chance, that is to discover a significant correlation purely by chance, we considered a relationship to be significant if the p-value was lower than 0.01. We also provided the t-value of each factor, which provides insight on the direction (positive or negative) and magnitude of the effect. The results of the multiple regression (see Table 7.8) indicate that %Protected, dominance and %Unstable AV vendor have a statistically significant ( $p < 0.01$ ) relationship with the infection rate. While %Protected is found to have a negative association with the infection rate, dominance and %Unstable AV vendor have a positive relationship.

Table 7.8 Multiple general linear regression (N=126 countries)

Indicator	$\beta$	Std. Error	t-value	p-value
%Protected	-0.33	0.08	-4.18	5.50e-05*
Richness	-0.09	0.07	-1.25	2.14e-01
Concentration	0.26	0.13	1.86	6.46e-02
Dominance	0.39	0.14	2.76	6.73e-03*
%Unstable AV vendor	0.41	0.08	4.98	2.19e-06*
R <sup>2</sup> adjusted	0.53			
F-statistic	29.25			
Degree of freedom	5			
Df (residuals)	120			
p-value	2.57e-19			

As the infection rates are function of the protection status (see Figure 7.3), we investigated to see if our previous findings apply to users when stratified by protection status. We classified users as being protected if they had an enabled AV during the entire study and unprotected if they had either an out of date, expired, snoozed or off AV, or no AV installed. Protected users got an average infection rate of 1.33% (SD=0.0067, 95% CI=0.0059-0.0076) while unprotected users got an average infection rate of 21.43% (SD=0.1315, 95% CI=0.1171-0.1501). As a comparison, the average infection rate for all users having an AV installed, regardless of the status of the latter, was 2.02% (SD=0.0124, 95% CI=0.0110-0.0141).

The Pearson correlation coefficients were first computed to identify any potential statistical association between the indicators and the infection rates by protection status. From Table 7.9 we can see that the relationships do not differ between protected and unprotected users. Moreover, the correlations found are similar to our previous findings (see Table 7.6): high %Protected and richness are associated with lower infection rates and high instability is associated with higher infection rates.

Table 7.9 Pearson correlation coefficients between infection rates and indicators by protection status (N=126)

	Protected		Unprotected	
	<i>r</i>	p-value	<i>r</i>	p-value
%Protected	-0.51	1.08e-09*	-0.06	9.97e-14*
Richness	-0.40	2.63e-06*	-0.40	4.24e-06*
Concentration	0.01	9.73e-01	-0.06	5.19e-01
Dominance	0.05	5.71e-01	0.13	1.45e-01
%Unstable AV status	0.60	6.54e-14*	0.86	0.00e-01*
%Unstable AV vendor	0.59	4.29e-13*	0.84	0.00e-01*

\*Statistically significant at 0.05 level

To better estimate the unique effect of each indicator, we performed a multiple general linear regression for each protection status (e.g. protected and unprotected). It appears the main difference between protected and unprotected users is the effect of dominance (see Table 7.10). While dominance is not related to infection rates for protected users, a negative significant association is found for unprotected users, meaning that higher dominance is associated with higher infection rates for unprotected users but not for protected.

Table 7.10 Multiple general linear regression by protection status (N=126 countries)

Indicator	Protected				Unprotected			
	$\beta$	Std. Error	t-value	p-value	$\beta$	Std. Error	t-value	p-value
%P	-0.29	0.07	-4.04	9.62e-05*	-0.19	0.06	-3.13	2.19e-03*
R	0.01	0.06	0.24	8.13e-01	0.04	0.05	0.69	4.91e-01
C	0.22	0.12	1.82	7.03e-02	0.22	0.11	2.14	3.43e-02
D	0.31	0.12	2.52	1.32e-02	0.30	0.11	2.82	5.53e-03*
%UAVV	0.58	0.07	7.82	2.34e-12*	0.73	0.06	11.38	0.00e-01*
R <sup>2</sup> adjusted	0.62				R <sup>2</sup> adjusted	0.72		
F-statistic	42.20				F-statistic	65.85		
Df	5				Df	5		
Df (residuals)	120				Df (residuals)	120		
p-value	0.00e-01				p-value	0.00e-01		

#### 7.4.1 Activity

The country level analysis allowed to identify a statistically significant correlation between the proportion of computers running an enabled AV product and malware infection rates. Intuitively, as the proportion of users running an enabled AV product increases, the rate of malware infections among users that have an AV product installed decreases. However,

what is less intuitive, is that the infection rates for unprotected users also tend to be lower in countries with higher proportion of users protected.

A first explanation for this is that unprotected users benefit from a herd immunity effect from protected users. This explanation can be explored by examining the correlation between the protection coverage –the proportion of protected users among all users– and the infection rates for unprotected users. A strong positive relationship ( $r=-0.85$ ,  $p\text{-value}=0.00e-01$ ,  $N=127$ ) was found between the protection coverage and the infection rates for unprotected. As shown in Figure 7.4, higher protection coverage is associated with lower infection rates for unprotected users. Although this broad correlation may provide empirical evidence of a herd immunity effect, proper validation should be achieved by conducting further studies designed for the purpose.

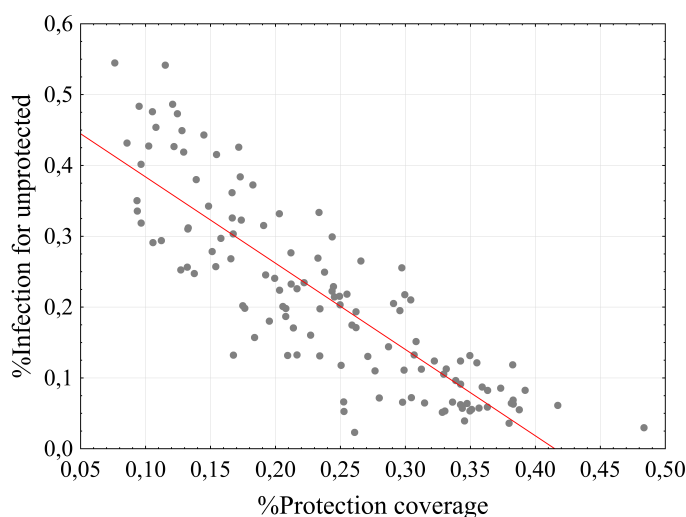


Figure 7.4 Infection rates for unprotected function of the protection coverage

A second possibility is that the proportion of users running an enabled AV product acts as a country level marker for users investment in the security of their computers. This would indicate that users in countries with higher protection coverage are less infected because they tend to be more aware of security risks and less likely to engage in risky behaviours. The validation of this explanation would require to conduct either country level studies based on aggregated measures of user security awareness or large-scale user studies.

**Findings:** Among users that have an AV product installed, 10% are not actively protected with up to date signatures.

Higher AV activity is significantly associated with lower infection rates, regardless of the protection status.

### 7.4.2 Diversity

Measures of diversity in terms of richness and dominance were significantly correlated with the infection rates. The degree of concentration, however, was not found to be significant in either of the regression models.

Richness was significantly and negatively correlated with infection rates of all users, whether protected or not. However, the correlation was not significant in the regression model after controlling for the other factors. One explanation could be that countries with higher AV adoption are more likely to have a diversified AV market.

A positive significant correlation was found between dominance and infection rates based on the regression models for all users and unprotected users, but not for protected users. One potential explanation could be that users in countries with AV monoculture are more vulnerable. The term monoculture refers originally to an agricultural practice of producing or growing one single crop over a broad area for several consecutive years. Since all plants are genetically similar, they are more vulnerable and less resistant to infections by pathogen, insects, or environmental conditions. If a new disease strikes to which they have no resistance, the entire population of crops can be destroyed. If we extend the principle to AV ecosystems, an AV monoculture would occur when the AV market is dominated by one single vendor. Therefore, population in such ecosystem are more likely to be infected when exposed to a new malware against which the AV product is not able to protect. This hypothesis should, however, be confirmed with theoretical models and validated with proper experiments in order to better understand the nature of the statistical association.

**Findings:** The global AV ecosystem has a degree of concentration around 0.92, meaning that it is highly diversified.

Higher AV dominance is significantly correlated with higher infection rates for all users that have an AV installed, as well as for unprotected users.

### 7.4.3 Stability

All indicators of AV stability were negatively significantly correlated with infection rates. These statistical associations were also found to be significant from the regression models for all users, as well as for protected and unprotected users. Moreover, the changes in AV vendors appeared to be the indicator with the strongest effect (based on its t-value) on infection rates.

One plausible explanation is that users switched vendors because they got infected. To examine this explanation, we computed the rate of changes in AV vendors for users that did not get any infection over the 4 months. Overall, 8.20% of those cleaned users changed



AV vendor, while the rate for all users was 33.57%. Changes in AV vendors could therefore be a consequence of detections by MSRT and be interpreted as a potential marker of users' satisfaction regarding the protection provided by AV systems. From those findings, AV vendors should make sure to detect the malware families covered by MSRT if they want to retain their customers.

**Findings:** Over the studied period, 40.47% of users changed their AV status and 33.57% changed AV vendor.

Higher stability, both in AV status and AV vendor, is significantly associated with lower infection rates for all users, whether having an AV product installed or not.

## 7.5 Discussion

This study and its results are subject to a number of limitations. First, there is an inherent bias to our results because our sample population is drawn from Windows systems running MSRT; thus, it does not provide insight into Windows systems that do not run Windows Update, and it does not give insight into the performance of AV on non-Windows systems such as MacOS and Android. However, given that there are more than one billion computers regularly running MSRT, patterns discovered in this population are important on their own, whether or not they are representative of patterns in other computational contexts.

Another significant limitation is that the infection rates as determined by MSRT are only for a subset of malware families. While these families may represent some of the most significant malware families on Windows, they are not a representative sample and so MSRT reported infection rates will be different from the overall malware infection rate. Nevertheless, given the significance of MSRT-targeted malware these infection rates are also of inherent interest.

This study was intended to be exploratory and not confirmatory, as our purpose was to develop indicators and investigate how they may relate to users' protection. Although ecosystem health measures cannot give predictive descriptions or identify causal mechanisms, they do provide case-by-case evaluations in real-world settings (Wilcox, 2001). Further studies should be conducted in order to validate our findings and investigate the nature of the associations we found.

Our results regarding the diversity of AV products used in practice roughly agree with other assessments of product market share (OPSWAT, 2014). Perhaps the best news coming out of this study is that almost 90% of the observed Windows systems are protected by anti-malware software and almost 90% of those systems are actively scanning with up to date signatures. While these numbers mean that the vast majority of users are maintaining their

systems properly, given the large number of Windows installations, these numbers also mean that millions of systems are not adequately protected.

There are also some clear indicators of ecosystem-wide dysfunction. Approximately one third of the systems running an AV product at the start of the study were using a different AV by the end of the four months. That is a remarkably high level of user churn; further, this churn is broadly distributed given the richness, high concentration, and low dominance of AV products in our sample. One hypothesis for this churn is that users are unsatisfied with AV products in general; clearly, though, this hypothesis requires further evaluation.

The country level analysis suggests the potential value of diversity in AV products, with countries with higher AV dominance having higher infection rates. The evidence that higher protection rates are correlated with lower infection rates in unprotected computers, suggests that mechanisms beyond the actual protection provided by AV products have protective effects, or that AV products provide protection for more than the host they are installed upon. Differentiating between these two effects may be a interesting area for future research as this question provides insight into how systems are or are not compromised by malware.

As AV products are continuously evolving, the process of evaluating the health of AV ecosystems should also evolve. In particular, while we hypothesize there is inherent value to diversity in the AV ecosystem and our work provides support for this hypothesis, it may be worth developing non-diversity based measures of ecosystem health in order to capture other important AV ecosystem-related health patterns. Further, a business ecosystem approach could be applied to evaluate AV ecosystem health defined by AV vendor, rather than by geographical unit. Such analysis could be seen as a complementary AV test to help customers choose an AV vendor based on specific indicators like users' loyalty, growth, or market share. Having said that, we believe the insights reported here show the potential benefits of our ecosystem health approach to studying AV performance.

## 7.6 Conclusion

In this paper we present a definition of antivirus ecosystem health based on a population's characteristic levels of activity, diversity, and stability. Using four months of telemetry data from MSRT, we calculated these health measures for a sample of more than one billion MSRT users and correlated them with MSRT —reported infection rates, in aggregate and on a per-country basis. Lowered infection rates were positively correlated with higher rates of AV activity, stable AV product usage and status, and AV product diversity. Higher AV activity also seems to be positively correlated with lowered infection rates on systems not protected

by AV software. While the results of this study cannot be considered definitive, they suggest that further work into measures of AV ecosystem health may produce significant insights into the performance of antivirus systems in practice.

**Acknowledgements** The authors would like to thank the Microsoft Malware Protection Center (MMPC) for granting us access to the MSRT telemetry data and for supporting this work.

## CHAPITRE 8    ARTICLE 5 : ARE THEY REAL ? REAL-LIFE COMPARATIVE TESTS OF ANTI-VIRUS PRODUCTS

*Published in the Proceedings of the 26th International Virus Bulletin Conference 2016.*

**Authors**    Fanny Lalonde Lévesque<sup>1</sup>, José M. Fernandez<sup>1</sup>, Dennis Batchelder<sup>2 3</sup>, Glaucia Young<sup>3</sup>

**Institutions**    École Polytechnique de Montréal<sup>1</sup>, AppEsteem<sup>2</sup>, Microsoft Corporation<sup>3</sup>

**Abstract**    This paper presents a novel methodology for conducting anti-virus (AV) tests based on real-life usage. In such tests, AV products are evaluated through long-term field studies where actual customers use the products in environments of their choice.

Using data collected from the Microsoft Malicious Software Removal Tool (MSRT) and Microsoft's Windows Defender on millions of systems, we conduct a large-scale comparative test of AV products. We describe our experimental design and present the results of the first test of this kind, aimed at evaluating AV products under real-life scenarios rather than in a controlled environment. Our findings show that AV performance varies significantly as a function of external factors that include user factors, environmental factors, and malware types.

### 8.1 Introduction

Typical anti-virus (AV) evaluation methods are based on automated tests performed in controlled environments. While these tests are adequate to evaluate the efficacy of AV products under specific scenarios, they do not measure the field efficacy of AV products as deployed on machines operated by real users. As many malware infections rely on direct or indirect user action, these situations cannot accurately be reflected in lab settings or theoretical models. Moreover, users may also impact the ability of AV products to prevent malware infection. For example, ignoring the dialog boxes or misconfiguring the product might result in a compromised system.

Given that users are involved in both the infection and protection process, it thus seems natural to adopt a *human-in-the-loop* approach to testing AV products. By including the user in the evaluation process, one could gain a better understanding of how the interactions

between users, malware and AV products can influence AV performance in the field. In other words, the protection offered by AV products can be affected by external factors beyond the quality of the engine.

Thus, to address both the questions of how to evaluate AV in real-life settings and how AV performance varies as a function of external factors, we developed a novel methodology to assess the ability of AV products to prevent malware infections in the field. This approach involves conducting longitudinal observational studies where deployed systems are assessed and compared based on their rate of malware infections.

In this paper, we report on a large-scale, real-life comparative test of AV products. The study involves more than 26 million Windows 10 systems that are assessed for a period of four months. Malware infections are computed using large-scale telemetry data from the MSRT and Windows Defender. We conduct initial testing to investigate how some external factors, such as malware types, user factors and environmental factors, affect the effectiveness of AV products as used in real-life settings. The remainder of the paper is organized as follows: Section 8.2 details the concept of real-life anti-virus evaluation. In Section 8.3 we describe our study design and present the results in Section 8.4. In Section 8.5, we discuss the benefits and limitations of this type of AV test. We conclude and discuss future work in Section 8.6.

## 8.2 Real-life anti-virus evaluation

Much as new drugs or medical interventions are studied first in the lab and later in the field, AV testing could adopt a similar clinical approach. In vaccine development, for example, efficacy studies are used to evaluate how well a vaccine performs under optimal clinical conditions. Once the vaccine has been proved to be efficient, effectiveness studies, also known as field efficacy studies, are used to measure direct and indirect vaccine protection under real-life conditions. Since those conditions are frequently suboptimal compared with clinical conditions, vaccine effectiveness is often lower than in the efficacy studies. Yet, field efficacy studies are needed to assess how real-life vaccine protection is affected by external factors, such as virus factors, host factors, storage, administration, availability and manufacturing of the vaccine.

Similarly, AV products could first be evaluated under controlled conditions. Further, effectiveness studies could be conducted as a complementary approach to in-lab evaluations. In such studies, AV products would be assessed over time on deployed systems used in real-life conditions. This could help better understand how AV products perform in actual use, and how external factors, such as the environment, the system configuration and user behaviour,

affect AV performance. Field studies of AV products could also provide crucial information to AV vendors on which aspect(s) of the product (user interface, detection, signature file updates, etc.) could be improved.

One potential way to assess AV protection in the field is to conduct computer security clinical trials, as proposed in 2009 by Somayaji *et al.* (2009). With such clinical trials, security software is installed and monitored over time on systems used in real-life settings. To prove the feasibility of this approach, a first pilot study aimed at evaluating one AV product was conducted at the École Polytechnique de Montréal in 2011 (Lalonde Lévesque *et al.*, 2012a,b; Lalonde Lévesque *et al.*, 2013). The study involved 50 participants, whose computers were instrumented and monitored for potential malware infections over a four-month period. While this study was limited to a single AV product, the methodology can also be adapted to conduct comparative clinical trials of AV products (Lalonde Lévesque *et al.*, 2012b).

Another suitable approach is to conduct observational studies of AV products. In contrast to experimental studies, such as clinical trials, AV products are not installed on systems. Rather, systems are monitored with their actual protection without any intervention. For example, Blackbird et Pfeifer (2013b) used MSRT data from millions of systems to evaluate how AV protection state impacts infection rates. Closer to our research is the work done by Lalonde Lévesque *et al.* (2015), where the authors also used MSRT data to measure the overall performance of the AV ecosystem over a four-month period. However, to the best of our knowledge, there has been no *real-life comparative evaluation* of AV products published in the literature. The key contribution of this work is therefore the realization of the first comparative test of this kind based on large-scale security telemetry data.

### 8.3 Study design and methods

By conducting this first test, we wanted to: 1) develop and test the validity and viability of a novel methodology for real-life comparative evaluation of AV products; and 2) determine how external factors, such as malware types, user factors and environmental factors affect AV performance.

#### 8.3.1 Cohort study design

We designed a cohort study to evaluate how AV products perform in real-life settings. This type of longitudinal observational study is often used in medicine, ecology, psychology, and social science to determine whether there is an association between an exposure to a risk (or

protective) factor and a disease. A cohort of individuals who are exposed to a specific factor and a similar cohort of individuals who are not exposed to the factor are followed over time until the disease of interest occurs.

The cohorts are then compared based on their respective frequency of disease, as presented in Table 8.1. A is the number of exposed individuals who developed the disease, B is the number of exposed individuals who did not develop the disease, C is the number of non-exposed individuals who developed the disease, and D is the number of non-exposed individuals who did not develop the disease.

Table 8.1 Frequency of disease by cohort

	Develop disease	Do not develop disease
Exposed	A	B
Not exposed	C	D

From Table 8.1, the relative risk (RR), that is the ratio of the incidence rate between the exposed and the non-exposed cohorts, can be calculated as follows:

$$RR = \frac{A/(A+B)}{C/(C+D)} \quad (8.1)$$

The confidence interval (CI) in which the true value of the RR is likely to be must be taken into account when interpreting the RR. An RR larger than 1 indicates that the exposure is a risk factor—the risk of developing the disease is higher for the exposed group. On the other hand, a RR smaller than 1 means that the exposure is a protective factor. And if the RR is equal to 1, this means that both groups (exposed and non-exposed) had the same ratio incidence. Hence, nothing can be said on the association between the exposure and the risk of developing the disease. Similarly, if 1 is included in the CI, meaning there is a chance that the RR is equal to 1, nothing can be said about the nature of the association. The CI is also important when comparing multiple results. If two CIs do not overlap, there is a statistically significant difference between the results. However, the opposite is not necessarily true. CIs may overlap, and yet there may be a statistically significant difference between the results.

From there, the effectiveness—the reduction in disease development between the exposed and non-exposed cohort— can be calculated from the RR:

$$\text{Effectiveness} = (1 - RR) \quad (8.2)$$

### 8.3.2 Study population

To conduct our cohort study, we used Windows 10 systems as our *study population*, malware infection as our *outcome* of interest, and being protected by a third-party AV product as our *exposure* factor (see 8.1). The exposed cohort (protected group) was defined as systems having a third-party AV product installed. As Microsoft’s Windows Defender is enabled by default on Windows 10 systems, we were not able to use “unprotected” systems as our non-exposed cohort. Instead, we used systems protected by Microsoft’s Windows Defender as our non-exposed cohort (comparison group). In other words, malware encounter (malware attacks blocked by Windows Defender) was used as a proxy for malware infection (successful malware attacks) for the comparison group. The study lasted four months, during which systems were assessed for potential malware infection.

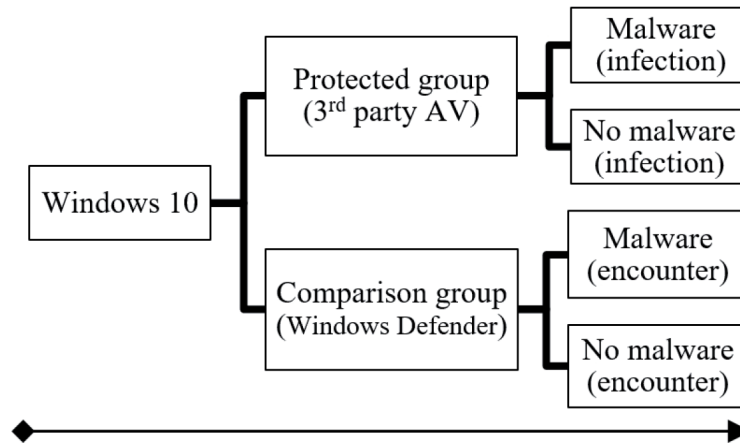


Figure 8.1 Cohort study design

### 8.3.3 Data collection

Data was collected from November 2015 to February 2016 on Windows 10 systems that reported monthly for the entire study period. All types of devices were included in the study (for example, desktop PCs, notebooks, and tablets) except mobile devices.

Information on those systems was coupled with demographic data from Microsoft Account, a single sign-on web service that allows users to log into various services provided by Microsoft (for example, Outlook, Skype, OneDrive). For each account, associated gender and age group were used. Gender could be male, female or unknown, and age was grouped into six categories (0–17, 18–24, 25–34, 35–49 and 50+). Accounts that had unknown age or gender or more than two devices associated with them were excluded from the analysis. Moreover, devices



that had more than one single account associated were also excluded. Internet Protocol (IP) geolocation was used to identify the country and the region of the systems. Regions were grouped into the following six categories: Africa & Middle East, Asia & Pacific, Australia, South & Central America, North America, and Europe. We further classified systems based on the 2015 Human Development Index (HDI) (United Nations Development Programme, 2015a) associated with their country of location. HDIs were grouped into the following categories: Very high ( $\text{HDI} \geq 0.8$ ), High ( $0.8 > \text{HDI} \geq 0.7$ ), Medium ( $0.7 > \text{HDI} \geq 0.55$ ) and Low ( $0.55 > \text{HDI}$ ).

### 8.3.4 Exposure (AV protection)

The exposure of a system (i.e. whether it belongs in the protected or comparison group) was determined from MSRT. MSRT is a malware cleaner utility that scans Windows computers for infections of specific, prevalent malicious software and helps remove these infections. MSRT is delivered and runs every month on more than one billion machines through Windows Update, as well as being available as a separate download from Microsoft. Upon execution, MSRT also calls the Windows Security Center (WSC) API to record information on the protection state of the system, such as the AV installed on the machine and its protection status (for example, enabled, expired, snoozed). As previously described in Section 8.3.2, systems protected by Microsoft's Windows Defender were included in the comparison group, and systems protected by third-party AV were included in the protected group.

In total, 26,956,360 Windows 10 systems were included in the study, with 16,464,730 in the protected group and 10,491,630 in the comparison group. Table 8.2 presents the basic characteristics of each group by user and environmental factors.

### 8.3.5 Statistical analysis

AV performance was determined from the rate of malware infection for each group (see Table 8.3). A is the number of systems protected by third-party AV that were infected by malware, B is the number of systems protected by third-party AV that were not infected by malware, C is the number of systems protected by Microsoft's Windows Defender that encountered malware, and D is the number of systems protected by Microsoft's Windows Defender that did not encounter malware.

Based on Table 8.3, the RR—the ratio of the incidence rate between the protected group and the comparison group— was computed from equation (8.2) with a CI of 95%. This means that there is a 95% chance the true RR value is included in the confidence interval. The AV

effectiveness (AVE) was then calculated as one minus the relative risk of malware infection (1-RR). This represents the proportionate reduction in malware infection rate between the protected group and the comparison group.

Table 8.2 Descriptive statistics by group

Factors	Protected group	Comparison group
Gender		
Female	35.90%	35.02%
Male	64.10%	64.98%
Age group		
0-17	4.57%	5.73%
18-24	18.16%	21.04%
25-34	20.70%	24.24%
35-49	25.55%	25.04%
50+	31.02%	23.95%
Region		
Africa & Middle East	1.76%	2.77%
Asia & Pacific	11.62%	11.24%
Australia	2.54%	2.31%
South & Central America	7.55%	6.95%
North America	39.54%	43.73%
Europe	36.99%	33.00%
HDI		
Very high	81.63%	79.73%
High	15.98%	15.69%
Medium	2.14%	3.95%
Low	0.24%	0.63%

Primary analysis of the overall AV effectiveness, that is the effectiveness of all AVs as a whole, was conducted first. A comparative analysis of AV effectiveness by vendor was further performed by malware types, gender, age group, region, and HDI category.

Table 8.3 Frequency of malware infection by group

	Malware	No malware
Protected group (3 <sup>rd</sup> party AV)	A	B
Comparison group (Defender)	C	D

## 8.4 Results

### 8.4.1 Anti-virus effectiveness

A total of 26,956,360 systems were assessed for malware infections. Among all 16,464,730 systems that were protected with a third-party AV (protected group), 1.22% were infected by malware during the study. And of the 10,491,630 systems in the comparison group, that is systems protected by Microsoft's Windows Defender, 1,568,122 encountered malware over the four-month period. In other words, if no AV was protecting the system, we could assume that 14.95% of those systems could have been infected by malware.

From this, we computed the relative risk and the overall AVE using equation (8.1) and (8.2) respectively:

$$RR = \frac{201,517/16,464,730}{1,568,122/10,491,630} = 0.0819$$

$$AVE = (1 - 0.0819) \times 100 = 91.81\%$$

The protected group was found to be less likely to have experienced any malware infection over the study period ( $RR = 0.0819$ ;  $CI\ 95\% = 0.0815-0.0822$ ). That is, compared with the systems with no AV installed (the comparison group), systems protected with a third-party AV product (protected group) had 0.0819 the risk of being infected by malware. Overall effectiveness of AV products in preventing malware infections was then estimated at 91.81% for the four-month period ( $AVE=91.81\%$ ;  $CI\ 95\%=91.77\%-91.85\%$ ).

We also investigated how the overall AV effectiveness varies as a function of AV protection status, types of malware, user factors, and environmental factors. Results in Table 8.4 show the estimates of the overall AVE and the 95% CI by factors considered. As previously mentioned, this means that there is a 95% chance the true value of AVE is captured within the confidence interval.

### AV protection status

Systems that had an enabled AV for the entire study were considered as having full protection (91.67%), and the other systems were considered having partial protection (8.33%). As expected, systems with full protection performed better ( $AVE = 91.97\%$ ) than systems with partial protection ( $AVE = 89.97\%$ ); though the difference was smaller than 2%. This is in line with previous studies (Blackbird et Pfeifer, 2013a; Lalonde Lévesque *et al.*, 2015) that

Table 8.4 Estimates of overall AVE at 95%

Factors		AVE	(95% CI)
AV protection status			
	Full	91.93	(91.89-91.97)
	Partial	89.80	(89.63-89.97)
Malware types			
	Malicious software	99.47	(99.46-99.48)
	Unwanted software	56.39	(56.14-56.65)
Gender			
	Female	89.39	(89.30-89.48)
	Male	92.54	(92.50-92.58)
Age group			
	0-17	87.65	(87.44-87.85)
	18-24	91.94	(91.86-92.01)
	25-34	92.27	(92.20-92.35)
	35-49	91.25	(91.16-91.33)
	50+	90.80	(90.70-90.90)
Region			
	Africa & Middle East	92.09	(91.91-92.27)
	Asia & Pacific	96.17	(96.11-96.22)
	Australia	88.52	(88.18-88.84)
	South & Central America	93.29	(93.20-93.38)
	North America	87.91	(87.81-88.00)
	Europe	91.76	(91.70-91.83)
HDI			
	Very high	88.72	(88.66-88.78)
	High	95.44	(95.40-95.49)
	Medium	92.64	(92.50-92.79)
	Low	94.51	(94.18-94.82)

found lower infection rates for systems protected with an enabled AV in comparison to other AV statuses (for example, out-of-date, expired, or snoozed).

### Malware types

As indicated in Table 8.4, we classified malware into two types: malicious software and unwanted software. Malicious software included viruses, trojans, hacks, rootkits, rogues, ransomware, infostealers, botnets, and exploits. Browser modifiers and adware were grouped into unwanted software. During the study, 0.08% of the systems in the protected group were infected with malicious software, and 1.15% were infected with unwanted software. Of the systems in the comparison group, 14.96% encountered malicious software, and 2.63%

encountered unwanted software over the four-month period.

Overall AVE was found to differ significantly between the two malware types: 99.47% for malicious software and 56.39% for unwanted software (see Table 8.4). As unwanted software is not universally identified as malware, this result can be explained by classification differences between third-party AV vendors and MSRT.

The low AVE for unwanted software can also indicate poor AV performance against this type of malware. One potential explanation could be differences in delivery vectors between malicious software and unwanted software. For example, many pieces of unwanted software are installed with “partial” user consent, either through social engineering methods, or bundled and chained installs. Although the installation required user interaction, this does not mean the software actually gained “full” consent from the user. It may not disclose all the components that will be installed, or it may install itself in a way that is very difficult to reverse. Moreover, the software can also update itself to install malicious components at a later time. This kind of deception can be difficult to detect with behaviour-based AV signatures and since the “vulnerability” is not in the operating system software itself (but is rather exploitation of the user), it cannot be mitigated with software updates.

## User factors

To explore how AV performance relates to user factors, we estimated the overall AVE by gender and age groups (see Table 8.4). A statistically significant difference in overall AVE was found between genders: effectiveness was estimated at 92.54% for male and 89.39% for female. Overall AVE by age groups (see Table 8.4) ranged from 87.65% to 92.27%; it increases until 25–34, after which it decreases. The 0–17 age group had the lowest AVE (87.65%), followed by the 50+ age group (90.80%), and the 25–34 age group had the highest AVE (92.27%). Interestingly, AV performance was found to differ significantly between all age groups. As the CIs do not overlap, we can assume that the differences in overall AVE are statistically significant.

Those gender and age variations in effectiveness could be explained by differences in the AV protection status. This would imply that females, young users (0–17), and users over 50 could be more likely to have partial protection, reducing the ability of the AV to protect against malware. To partially test this hypothesis, we computed for each user factor the percentage of systems that had partial protection over the study. However, the results in Table 8.5 suggest that AV protection status may not account for gender and age differences in AV performance.

Table 8.5 AV protection status by user factor

Factors	Partial protection
Gender	
Female	3.83%
Male	4.5%
Age group	
0-17	4.56%
18-24	4.83%
25-34	4.40%
35-49	4.00%
50+	3.81

Another plausible explanation could be differences in malware exposure and user behaviour when faced with malware attacks. For example, some users (female, 0–17, 50+) could be more vulnerable to sophisticated malware attacks that exploit psychological and social engineering techniques, such as deception, manipulation, or intimidation, to infect the victim’s system and circumvent AV detection. This can be investigated by looking at the infections for ransomware and rogue malware —two types of malware that use deceptive fear to trick users into downloading malicious software (a trial version of a bogus security program or a fake software update).

Interestingly, we found the 50+ age group to be the most infected with these types of malware: they represented 41.34% and 75.00% of the infections for ransomware and rogue malware, respectively. Similarly, females and users in the 0–17 age group could be more susceptible to other delivery vectors based on social engineering (for example, phishing or spearphishing, social networking attacks, and clickbait). This is consistent with previous studies that found evidence that females (Sheng *et al.*, 2010; Jagatic *et al.*, 2007) and young users (Sheng *et al.*, 2010; Kumaraguru *et al.*, 2009) are more vulnerable to phishing and spear phishing attacks.

While our findings indicate that some user groups (female, 0–17, 50+) are more likely to be infected with malware, they do not explain why these users were more infected even though they had an AV product installed. A more fine-grained analysis is needed to better understand the underlying human and technical mechanisms behind those gender and age differences in AV performance.

### Environmental factors

We classified systems by region and HDI based on their country of location. As presented in Table 8.4, North America (AVE=87.91%) and Australia (AVE=88.52%) had the lowest

overall AVE, and Asia had the highest effectiveness (AVE=96.17%). When looking at the overall AVE by HDI, we found the HDI category “Very high” to have the lowest value (AVE=88.72%), and the HDI category “High” to have the highest value (AVE=95.44%). The other HDI categories (medium and low) had AVEs between 92.64% and 94.51%, respectively. All AVEs by region and HDI were found statistically to be significantly different.

One possible explanation for these environmental variations in overall AVE could be demographic differences between user populations. To explore this hypothesis, we computed for each HDI category the prevalence of each user factor (age group, gender) in the protected group.

Table 8.6 User factors by HDI

Factors		Very high	High	Medium	Low
Gender					
	Female	38.27%	26.19%	19.91%	19.99%
	Male	61.73%	73.81%	80.09%	80.01%
Age group					
	0-17	4.33%	5.90%	4.05%	3.014.33%
	18-24	15.73%	29.10%	28.98%	21.89%
	25-34	18.16%	31.83%	32.70%	35.18%
	35-49	26.19%	22.62%	22.98%	27.01%
	50+	35.59%	10.55%	11.29%	12.91%

From Table 8.6, the HDI category “Very high” has the highest percentage of females and users in the 50+ age group. Based on our previous results on user factors, those demographic differences may explain, to some extent, the low AVE (87.91% for systems located in countries with very high HDI).

These environmental variations may also be explained by *geographical* differences in the malware landscape. For example, financial malware, such as ransomware and banking malware, are more prevalent in wealthier countries (Savage, Kevin and Coogan, Peter and Lau, Hon, 2015). Conversely, other geo-malware will avoid infecting systems located in specific countries or with particular languages.

#### 8.4.2 Anti-virus effectiveness by vendors

To conduct a comparative analysis of effectiveness, we grouped systems by AV vendors based on the AV product installed. In other words, we created protected groups by AV vendors and compared each group to systems protected by Microsoft’s Windows Defender. For example, Figure 8.2 illustrates the design of a cohort study aimed at evaluating the effectiveness of

two AV vendors.

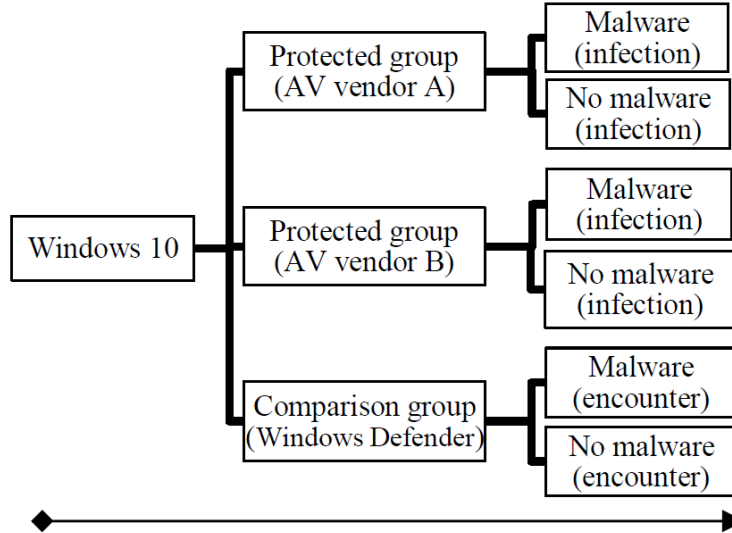


Figure 8.2 Comparative cohort study design

In this section, we report on the effectiveness of the 10 most prevalent AV vendors, which represents 90.74% of the systems protected by third-party AVs. Table 8.7 presents for each vendor the estimate of effectiveness and the 95% CI.

Table 8.7 Estimates of AVE at 95%

Vendor	AVE	(95% CI)
A	98.03	(97.95-98.11)
B	95.52	(95.37-95.66)
C	95.41	(95.29-95.53)
D	93.77	(93.59-93.95)
E	93.58	(93.43-93.73)
F	92.37	(92.27-92.47)
G	90.94	(90.82-91.05)
H	90.86	(90.69-91.03)
I	90.53	(90.44-90.62)
J	90.01	(89.92-90.11)

Results ranged from 90.01% to 98.03%. Most AVEs were found to be statistically significantly different between vendors; there was no overlap between the 95% CIs. However, we were not able to assume with certainty that the difference in AVE was significant for three of the 45 pairwise comparisons between vendors.



## Malware types

Overall, AV products performed better at preventing infections from malicious software than from unwanted software. AV vendors that performed better for malicious software also performed better for unwanted software.

Table 8.8 Estimates of AVE at 95% for different malware types

Vendor	Malicious software		Unwanted software	
	AVE	(95% CI)	AVE	(95% CI)
A	99.86	(99.84-99.88)	89.61	(89.15-90.05)
B	99.88	(99.85-99.90)	75.26	(74.42-76.07)
C	99.73	(99.70-99.75)	75.50	(74.80-76.17)
D	98.88	(98.80-98.95)	70.80	(69.86-71.71)
E	99.51	(99.47-99.55)	66.24	(65.40-67.06)
F	99.85	(99.84-99.86)	57.53	(56.98-58.08)
G	99.65	(99.63-99.67)	50.45	(49.79-51.09)
H	99.48	(99.44-99.52)	50.99	(50.03-51.93)
I	99.46	(99.44-99.48)	49.16	(48.63-49.69)
I	99.42	(99.40-99.45)	46.42	(45.85-46.99)

Interestingly, the ranking for malicious software is different from the ranking presented in Table 8.7. For example, vendor D has a significantly lower effectiveness—it moved from the fourth position (#4) in Table 8.7 to the last one (#10) in Table 8.8. Although there is only a 1% variation between the lowest AVE (98.88%) and the highest (99.88%), most differences are statistically significant.

Among the 45 pairwise comparisons between vendors, 36 were found to be significant. However, we were not able to assume with certainty that the remaining 12 pairwise comparisons were significant.

When comparing the ranking with Table 8.7 for unwanted software, only vendor B had a different position—it was permuted with vendor C. The difference between the highest (89.61%) and the lowest (46.42%) AVE is also much more important in comparison with the results for malicious software. Of the 45 pairwise comparisons of AVEs between vendors, 43 were found to be significant, meaning that AVE variations between vendors are much more important for unwanted software than for malicious software.

## User factors

All vendors were found to perform significantly better at protecting male users than female users (see Table 8.9). However, those gender differences in AVE were not constant across

vendors. For example, vendor A had a 0.64% difference while vendor D had a difference of 5.60% between male and female groups.

Effectiveness for male users ranged from 90.44% to 98.19%, while for females it ranged from 87.29% to 97.55%. In comparison with the results in Table 8.7, there was no variation in the ranking for male users. However, some vendors moved in the ranking when looking at female users. For example, vendor C moved from the third position (#3) to the second position (#2), and vendor D lost two positions, moving from (#4) to (#6). Of the 45 pairwise comparisons between vendors, 42 were confirmed to be significant for male and also for female groups.

Table 8.9 Estimates of AVE at 95% by gender

Vendor	Male		Female	
	AVE	(95% CI)	AVE	(95% CI)
A	98.19	(98.10-98.28)	97.55	(97.35-97.74)
B	96.14	(95.99-96.29)	93.53	(93.14-93.90)
C	95.24	(95.07-95.41)	94.57	(94.34-94.79)
D	94.92	(94.75-95.08)	89.32	(88.67-89.93)
E	94.22	(94.06-94.37)	91.95	(91.55-92.34)
F	93.08	(92.97-93.19)	89.91	(89.70-90.13)
G	91.72	(91.59-91.84)	88.33	(88.06-88.59)
H	92.01	(91.84-92.19)	87.23	(86.79-87.65)
I	91.54	(91.44-91.64)	87.29	(87.06-87.51)
I	90.44	(90.32-90.55)	87.82	(87.61-88.02)

Similarly to the overall AVE (see Table 8.7), all vendors except one were found to have lower effectiveness for the 0–17 age group (see Table H.1 in Appendix H). The exception was vendor D, where the lower effectiveness was for the 50+ age group. In contrast, the age group with the highest effectiveness was not constant across vendors.

Interestingly, some age groups had larger intervals between the lowest and the highest values. For example, AVE ranged from 85.01% to 97.57% for the 50+ age group, which represents a variation of 12.95%. This is in sharp contrast with other age groups, such as 18–24, 25–34, and 35–49, which had 7% intervals.

The ranking by vendor is also different for all age groups in comparison to the results in Table 8.7, though vendor A is always in first position. Most vendors only gained or lost one position. The exception was the 50+ age group, where vendor D moved from the fourth position (#4) to the last one (#10), and vendor E gained three positions. We also explored how results vary within vendors. Vendor D had significant differences in AVE for all age groups, as well as the higher variation (11.09%) between its minimum and maximum value. For the other vendors, the number of significant pairwise comparisons ranged from four to

nine out of 10.

Overall, those results suggest that AV performance is not constant across user factors. That is, some vendors performed better than others for specific gender and age groups. Furthermore, most of these gender and age differences in effectiveness were found to be significant between and within vendors.

As previously mentioned in Section 8.4.1, those age and gender variations in AV effectiveness may be attributed to differences in malware exposure and user behaviour when faced with malware attacks. Another plausible explanation could be user differences in attitude and behaviour towards AV products. This would imply that specific user groups, such as female, 0–17 and 50+, could be more likely to misconfigure the AV product, misunderstand the security warnings, or ignore the dialog boxes from the AV.

Due to high variance among user attitude, knowledge and risk aversion in security, AV products designed to suit the “average user” may not accommodate all user groups (Egelman et Peer, 2015). In other words, one AV product may not fit all.

### **Environmental factors**

Similar to the overall analysis results (see Table 8.7), North America had the lowest AVE for all vendors except for vendor B, which was less effective for Australia (see Table H.2 in Appendix H). North America was also the region with the highest variance between vendors, ranging from 80.88% to 97.45%.

Conversely to Table 8.7, Asia & Pacific was not always the region with the highest AVE. Some vendors were more effective for Asia & Pacific, while others performed better for Africa & Middle East, or South & Central America. Moreover, not all vendors had the same fluctuations between regions. For example, vendor A has a variation under 3%, while vendor D has a variation of almost 18%.

Rankings by region were also different from Table 8.7, though most vendors only lost or gained between one and three positions. The exceptions were vendors C and D, where C lost six positions for Europe, and D lost six positions for North America and Europe. Overall, most of the pairwise comparisons between and within vendors were significant.

The HDI category “Very high” was the category with the lowest AVE for all vendors (see Table H.3 in Appendix H). It was also the category with the highest variance in AVE, ranging from 81.10% to 97.31%. However, the HDI category with the highest effectiveness was not the same across vendors. Rankings were also different for each HDI category, though vendor A was always the first. Most vendors lost or gained between one and three positions in the

rankings. The exceptions once again were vendors C and D, where C lost six positions for the categories “Very high”, “Medium” and “Low”, and D lost seven positions for the category “High”. Finally, most of the pairwise comparisons in AVE between and within vendors were found to be significant.

## 8.5 Study limitations

While cohort studies allow us to determine whether an exposure is associated with an outcome, their results can be highly sensitive to bias and unknown external variables that may affect the results. This study and its findings are therefore subject to a number of limitations and potential bias.

First, our study population is drawn from Windows 10 machines only, it does not provide insight into other version of Windows (for example, Windows Mobile, Vista, XP, etc.). Furthermore, the analysis was limited to Windows 10 systems having a Microsoft Account with known age and gender information. Hence, our study population may not be representative of the entire population of Windows 10 systems.

Second, although our protected group and comparison group originate from the same population (Windows 10), we cannot ensure with certainty that they only differ by their exposure (third-party AV or Windows Defender). That is, customers of third-party AVs (protected group) may also differ in other factors, such as attitude and behaviour towards security, in comparison to users that are protected with Microsoft’s Windows Defender (comparison group). This would imply that other factors, beyond the exposure (AV protection), may affect the outcome (malware infection).

As we limited our outcome to malware families covered by MSRT, we only evaluated how third-party anti-virus products are effective at preventing malware infections for a subset of malware families. Although these families may represent some of the most prevalent threats on Windows, they do not cover targeted attacks and zero-day attacks.

Furthermore, MSRT-targeted malware may not reflect the priorities and viewpoint of all third-party AV vendors. This is particularly true in the context of unwanted software. While this might have introduced a selection bias, the latter was equally applied to all third-party AV vendors. Given the significance of the malware families covered by MSRT, these infections are of inherent interest, whether or not they are representative of the entire malware landscape on Windows 10.

Another significant limitation derives from how we assessed our outcome (malware encounters) for the comparison group. As the encounters reported depend on the ability of Mi-

Microsoft's Windows Defender, this may have led to an underestimation of malware encounters. On the other hand, malware encounters as a proxy for malware infections may result in an overestimation. As many pieces of malware target specific languages, countries, or software vulnerabilities, we cannot assume that those encounters would have been successful infections if no anti-virus was protecting the system. It would be interesting in future work to use telemetry data from other AV vendors to complement our data on malware encounters for the comparison group.

Although gender, age, region, and HDI were included in our analysis, we cannot ensure that our results are not affected by other unknown external factors that may affect the ability of anti-virus products to prevent malware infections. Moreover, our study only estimated the *non-adjusted* anti-virus effectiveness, meaning that we did not control for those external factors. It would be interesting for future work to compare the non-adjusted and *adjusted* effectiveness. That is, the effectiveness when the effect of other, external, factors is netted out.

Finally, we only estimated how third-party anti-virus vendors are effective at preventing malware infections for their customers. In other words, the same AV vendor may have a different effectiveness when protecting customers from other vendors. This implies that our ability to generalize our results to a more universal population of Windows 10 systems may be limited. One way to obtain more universal results would be to conduct computer security clinical trials where anti-virus products could be assigned to systems randomly, limiting any potential customer-based bias.

## 8.6 Conclusion

We have presented the design and the results of a first real-life comparative evaluation of anti-virus products. Overall, we found that AV effectiveness differs significantly by malware types and populations. We also explored the potential underlying causes between those differences in AV performance.

The key insights are that, in practice, AVs perform differently, and their performance varies by population. As we hypothesized, this may be due to biases in malware selection or product design. In the first case, this might either be due to targeting of specific populations by malicious actors, or more likely due to the link between user behaviour and the attraction techniques employed by the former.

On the other hand, product design features, such as user interface, might make certain AVs work better or worse for some individuals. This seems particularly important for young

(0–17) and female users. Perhaps this is like safety equipment in cars: when seat belts and airbags were not developed with those populations in mind we had higher rates of injuries and fatalities for those groups.

A similar dynamic may also be taking place with AV software. This latter possibility might offer a potential opportunity for the AV vendors to improve their real-life efficacy with real users.

Indeed, this paper and its findings have important implications for the anti-virus industry. For AV vendors, this kind of test could help: 1) understand how they perform in real-life scenarios, 2) identify which aspect(s) of the product could be improved, and 3) identify user groups for which they are more (or less) effective at preventing malware infections. We believe a better understanding of what works best in the field for specific user groups is the first step towards the development of successful *user-differentiated* AV software.

Real-life anti-virus evaluation might also provide AV testers a viable and complementary approach to in-lab tests. Due to the realism of the testing environment and the independence of the threat selection process, real-life tests could offer performance results that are less prone to controversy and ethical issues, such as the creation of malware samples (Anti-Malware Testing Standards Organization Inc., 2016). For example, testers could conduct a retrospective study to evaluate how effective AV products are at preventing zero-day or targeted attacks on deployed systems. Other real-life tests could also be designed to evaluate the true benefit between free and paid AV products, or how specific features truly contribute to reduce the risk of malware infection in the field.

This study may also have potential implications for customers. The variations in effectiveness for different users groups suggest it might be advantageous for users to choose AV that better protects their respective user group. Ultimately, however, we believe that causal factors are more likely to be related to user behaviour rather than demographic characteristics. Thus, it might prove more accurately and ultimately useful to characterize AV effectiveness in terms of user behaviour profiles (gamers, social networkers, etc.), as was previously suggested (PC Security Labs, 2013).

In summary, we hope this work demonstrates the merit of real-life anti-virus evaluation for both the AV industry and the scientific community. In future work, we intend to address some of the limitations we mentioned in Section 8.5. In addition, an important open question that this work poses but that remains unanswered is whether variations in AV effectiveness for different user groups are mostly caused by differences in malware exposure between these groups or due to how these users interact with the AV product. Determining causality for these differences in effectiveness is paramount in order for the AV industry to be able to

improve value for *all* of its customers.

**Acknowledgements** The authors would like to thank the Microsoft Malware Protection Center (MMPC) for granting us access to the Windows Defender and MSRT telemetry data and for supporting this work. The authors would also like to thank Anil Somayaji, François Labrèche and Holly Stewart for their useful comments and suggestions on the paper.

## CHAPITRE 9 DISCUSSION GÉNÉRALE

Ce chapitre se veut une discussion générale sur les travaux réalisés dans le cadre de la thèse, ainsi que sur les différentes contributions et implications des résultats obtenus.

### 9.1 Modèle de prévention des attaques par logiciels malveillants

*Développer et appliquer un modèle basé sur l'approche de la santé publique pour la prévention des attaques par logiciels malveillants.*

#### 9.1.1 Identification des déterminants

*Identifier les causes et les corrélats reliés aux attaques par logiciels malveillants.*

**Facteurs socio-environnementaux** *Quels sont les facteurs socio-environnementaux qui sont reliés au taux national d'infections par logiciels malveillants ?* Les résultats de notre analyse multi-pays présentée au Chapitre 4 suggèrent que l'éducation, l'économie, la technologie et la sécurité informatique sont tous des facteurs associés au taux national d'infections par logiciels malveillants. Fait intéressant, l'analyse montre que l'effet de ces facteurs peut varier en direction et magnitude selon le statut socio-économique des pays. Ceci implique, entre autres, que le contexte socio-économique d'un pays doit être considéré dans le développement et l'évaluation de politique en sécurité des systèmes d'information. En d'autres mots, une intervention prouvée comme étant efficace dans un pays pourrait ne pas avoir l'effet escompté dans un pays dont la réalité socio-économique est différente. De plus, ces résultats invitent à une priorisation des efforts, où les facteurs de protection les plus importants devraient être ciblés en premier.

**Facteurs comportementaux** *Quels sont les comportements des usagers qui sont reliés au risque d'attaques par logiciels malveillants ?* L'analyse des données comportementales au Chapitre 5 a permis d'identifier plusieurs corrélats significatifs. Notamment, le volume d'activité en ligne, la visite de certains types de sites Web, le téléchargement de fichiers exécutables à partir d'Internet, et le recours à des réseaux pair à pair ont été identifiés comme des facteurs de risque. En outre, ces résultats suggèrent que les comportements à *risque* ne sont pas limités à ce qui est traditionnellement associé à un risque plus élevé, tel que l'utilisation des réseaux pair à pair. Bien que certains de ces facteurs de risque peuvent sembler intuitifs,



l'identification ainsi que la quantification de leurs effets respectifs sont des éléments essentiels au développement de meilleures formations et stratégies de protection.

**Facteurs démographiques** *Quel est l'effet indépendant de l'âge et du genre sur le risque d'attaques par logiciels malveillants ?* Les résultats obtenus au Chapitre 6 lors de notre étude cas-témoins suggèrent que l'âge et le genre ont un effet indépendant significatif sur le risque d'attaques par logiciels malveillants. De plus, cet effet semble varier en direction et en magnitude selon le type de logiciels malveillants. La compréhension de ces relations est essentielle pour le développement d'interventions efficaces visant à prévenir ou réduire le risque d'attaques par logiciels malveillants. Par exemple, comprendre la cause de ces différences permettrait de mettre au point des stratégies différenciées par l'âge et le genre et d'adapter ces dernières selon le type de menaces. Similairement, une meilleure compréhension de l'impact de ces facteurs pourrait permettre à l'industrie de la *cyberassurance* de développer des modèles de risque basés sur des faits et des données probantes.

### 9.1.2 Évaluation de stratégie

*Évaluer l'efficacité réelle des solutions antivirus à prévenir et/ou réduire l'occurrence des attaques par logiciels malveillants.*

**Évaluation agrégée** *Quel est l'état de santé de l'écosystème des logiciels antivirus ?* L'analyse réalisée au Chapitre 7 a permis de mesurer l'activité de l'écosystème global des solutions antivirus et d'évaluer à 90% le taux de système Windows qui ont une solution antivirus à jour. Bien que la très grande majorité des systèmes étudiés étaient activement protégés, cela signifie que plusieurs millions de système ne sont pas adéquatement maintenus et protégés. Quant à la diversité, les résultats suggèrent une valeur potentielle à cette dernière. Fait intéressant, les écosystèmes à haute diversité ont présenté un plus faible taux d'infections par logiciels malveillants pour les systèmes non protégés. Cette relation pourrait, par exemple, indiquer la présence d'un phénomène d'immunité grégaire (ou immunité de communauté) au sein des écosystèmes de solutions antivirus. En d'autres mots, la protection offerte par une solution antivirus irait au-delà du système sur lequel elle est installée. L'écosystème global de solutions antivirus s'est révélé être très instable ; environ un tiers des systèmes protégés avaient une solution antivirus différente à la fin des quatre mois. Bien que notre analyse ne permette pas d'en connaître la cause exacte, cette grande variation pourrait indiquer une certaine insatisfaction de la part des usagers à l'égard des solutions antivirus, motivant ainsi ces derniers à changer de produit. En conclusion, les résultats de ce travail suggèrent que l'approche de

santé d'écosystème peut être appliquée et se révéler utile dans le contexte des solutions antivirus. Notamment, une telle approche permet de i) surveiller l'activité d'un écosystème, ii) identifier et prédire des zones sujettes à amélioration, et iii) évaluer la performance agrégée de solutions antivirus dans un contexte réel d'utilisation.

**Évaluation comparative** *Quel est l'impact de l'environnement sur la performance des logiciels antivirus ?* Le Chapitre 8 a permis de montrer que la performance des solutions antivirus varie significativement en fonction de plusieurs facteurs externes, tel que l'utilisateur, le contexte socio-économique, ou le type de menaces. Les travaux réalisés ont de plus plusieurs implications pour l'industrie antivirus. Notamment, les tests en conditions réelles pourraient aider les vendeurs antivirus à : i) comprendre comment leurs solutions performant dans la vraie vie, ii) identifier quels aspects des produits pourraient être améliorés, et iii) identifier les populations pour lesquels leurs produits sont plus (ou moins) efficaces. Au niveau des testeurs, ce type d'évaluation pourrait offrir une approche viable et complémentaire aux tests réalisés en laboratoire. Ces travaux ont aussi une implication potentielle pour les usagers. Par exemple, les résultats indiquent qu'il pourrait être plus avantageux pour les usagers de choisir une solution antivirus qui semble plus efficace à protéger le groupe d'utilisateurs auquel ils sont associés.

## 9.2 Contributions et implications

Nous présentons dans la section qui suit un portrait de nos contributions ainsi que de leurs implications potentielles en sécurité des systèmes d'information.

**Facteurs de risque et de protection** Nos différentes analyses ont permis d'identifier plusieurs facteurs de risque et facteurs de protection reliés au risque d'attaques par logiciels malveillants. Entre autres, nous avons étudié l'impact de facteurs au niveau de l'environnement et des politiques (éducation, économie, technologie, sécurité des systèmes d'information, écosystème des antivirus), de l'utilisateur (âge, genre, comportement), et du système (antivirus). Il en ressort que le risque d'attaques par logiciels malveillants dépend d'une combinaison de facteurs multi-niveaux. L'identification de ces facteurs peut ainsi servir d'évidence et de soutien au développement de stratégies basées sur des données probantes. Notamment, plusieurs des facteurs identifiés, tel que l'éducation et le développement technologique, peuvent être influencés par des politiques nationales en matière de sécurité des systèmes d'information.

**Impact du contexte** En complément de notre analyse des facteurs de risque et de protection, nous avons étudié comment l'impact de certains facteurs varie en direction et importance selon le contexte. Notamment, nous avons mis en évidence que l'effet des facteurs socio-environnementaux et démographiques est influencé respectivement par le statut socio-économique du pays et le type de logiciels malveillants. Ces résultats suggèrent une valeur ajoutée au développement de stratégies écologiques, c'est-à-dire des stratégies multi-niveaux qui prennent en compte l'impact du contexte. De plus, une compréhension de l'impact du contexte peut permettre de supporter les efforts de priorisation, où les stratégies étant les plus efficaces dans un contexte donnée peuvent être prioriser.

**Efficacité réelle** L'évaluation de l'efficacité réelle des solutions antivirus a permis de mettre en évidence que cette dernière est inférieure aux performances rapportées en conditions contrôlées. Autrement dit, l'efficacité réelle d'une stratégie est influencée par différents facteurs externes inhérent à son contexte d'implémentation. En ce sens, notre analyse est un premier pas vers l'identification des facteurs externes (statut socio-économique, type de logiciels malveillants, profil de l'utilisateur) qui influencent la performance des antivirus lorsque déployés dans un environnement réel. Ces résultats ont entre autres plusieurs implications pour le développement de stratégies de prévention. Notamment, ils contribuent à renforcer l'importance de considérer le contexte lors de l'implémentation d'une stratégie, et à favoriser le développement de stratégies personnalisées.

**Modèle de prévention** Notre contribution principale consiste au développement et à l'application d'un modèle basé sur l'approche de la santé publique pour la prévention des attaques par logiciels malveillants. À cet effet, la présente thèse se démarque des travaux antérieurs du fait que nous appliquons différentes méthodes et approches de santé publique à un cas réel, à savoir, la prévention des attaques par logiciels malveillants. Les résultats obtenus lors de nos différentes analyses contribuent de plus à supporter l'application d'une approche inspirée de la santé publique en sécurité des systèmes d'information. L'adoption d'une telle approche est, à notre avis, non seulement viable, mais nécessaire afin de développer une stratégie globale de prévention et de protection de la population contre les attaques par logiciels malveillants.

### 9.3 Sécurité des systèmes d'information publique

Nous espérons avoir démontré par le présent projet de recherche la valeur de l'approche de la santé publique appliquée en sécurité des systèmes d'information. Ayant fait ses preuves dans

le domaine de la santé, le recours à une telle approche est, à notre avis, viable et bénéfique pour l'adoption d'une stratégie globale en sécurité des systèmes d'information.

Qui plus est, l'exploration de l'analogie entre la santé publique et la sécurité permet d'entrevoir la possibilité de considérer cette dernière comme un bien public (Mulligan et Schneider, 2011). Similairement à la santé, la sécurité des systèmes d'information répond aux deux critères nécessaires : la non-rivalité et la non-exclusion (Samuelson, 1954). La non-rivalité signifie que la consommation du bien, la sécurité des systèmes d'information, par un individu ne prive pas un autre individu de le consommer. En d'autres mots, la consommation du bien par un individu n'affecte pas la quantité disponible pour les autres. Quant à la non-exclusion, elle désigne le fait qu'un individu ne peut être privé de consommer ce bien. Autrement dit, la consommation de la sécurité des systèmes d'information ne peut être individualisée.

Dans cette optique, la sécurité des systèmes d'information *publique* peut se définir comme suit :

*“La sécurité des systèmes d'information publique est la science et l'art d'améliorer la sécurité des systèmes, réseaux, et usagers, et de réduire les inégalités en matière de sécurité à travers des efforts coordonnés. ”*

À l'instar de la santé publique, une approche axée sur la sécurité des systèmes d'information publique se penche sur un large éventail de facteurs et de conditions reliés à la sécurité afin d'influer ces derniers. Elle tient notamment en compte le fait que des facteurs qui sont indépendants du système, tels que l'utilisateur et l'environnement socio-économique, exercent une incidence sur la sécurité des systèmes d'information.

## CHAPITRE 10 CONCLUSION

Dans ce dernier chapitre, nous présentons en un premier temps une synthèse des travaux réalisés ainsi que des résultats obtenus. En un second temps, nous discutons des travaux futurs et terminons avec la conclusion.

### 10.1 Synthèse des travaux

La section suivante fait état des différents travaux réalisés dans le cadre de la présente thèse. En particulier, nous exposons les travaux de recherche reliés à l'atteinte de notre objectif général, ainsi qu'à nos deux objectifs spécifiques.

#### 10.1.1 Modèle de prévention

*Développer et appliquer un modèle basé sur l'approche de la santé publique pour la prévention des attaques par logiciels malveillants.*

En s'inspirant du cadre de travail en santé publique (voir Figure 2.2), nous avons proposé dans le Chapitre 2 un cadre de travail similaire pour le contexte de la sécurité des systèmes d'information. En particulier, nous avons appliqué ce modèle au contexte de la prévention des attaques par logiciels malveillants. Le modèle développé a été appliqué à deux niveaux afin d'atteindre nos objectifs spécifiques, soit l'identification de déterminants, et l'évaluation d'une stratégie de prévention. En ce qui concerne l'identification de déterminants, nous avons en un premier temps présenté l'état actuel des connaissances à ce niveau, et développé un modèle écologique des déterminants (voir Figure 3.1). En un second temps, nous avons réalisé trois études observationnelles visant à étudier les déterminants au niveau de l'environnement et des politiques, et au niveau de l'utilisateur. Quant à l'évaluation de l'efficacité d'une méthode de prévention, soit les solutions antivirus, nous avons présenté l'état actuel des méthodes de tests, et exploré l'analogie avec l'évaluation de nouveaux médicaments et interventions médicales. À cet effet, nous avons réalisé deux études en conditions réelles. La première visant à évaluer l'efficacité agrégée des antivirus, et la seconde visant à réaliser un test comparatif, le tout dans un environnement réel.

#### 10.1.2 Identification des déterminants

*Identifier les causes et les corrélats reliés aux attaques par logiciels malveillants.*

**Facteurs socio-environnementaux** Nous avons, dans le Chapitre 4, réalisé une étude écologique multi-pays afin d’investiger l’impact de facteurs socio-environnementaux sur le taux national d’infections par logiciels malveillants. La variable dépendante, le taux national d’attaques par logiciels malveillants, a été évaluée à partir de la proportion de systèmes Windows non protégés par une solution antivirus qui ont été infectés au cours d’une période de quatre mois. Les informations sur les infections par logiciels malveillants ont été collectées mensuellement par MSRT sur plus de 10 millions de systèmes dans 186 pays. Quant aux facteurs socio-environnementaux, nous avons pris en compte différentes variables permettant de refléter la réalité socio-économique et technologique des pays, ainsi que leur posture en matière de sécurité des systèmes d’information. Cette étude a notamment permis de i) identifier plusieurs corrélats tels que l’éducation, le développement technologique, la performance économique, et la posture nationale en sécurité des systèmes d’information, et ii) mettre en évidence que l’effet des déterminants varie en direction et magnitude selon le statut socio-économique.

**Facteurs comportementaux** Dans le Chapitre 5, nous avons analysé des données comportementales provenant d’une étude utilisateurs réalisée par Lalonde Lévesque *et al.* (2013). En particulier, nous avons étudié le comportement de 50 usagers durant une période de quatre mois afin d’identifier les comportements des usagers étant corrélés avec un risque plus élevé d’attaques par logiciels malveillants. Nous avons notamment été en mesure d’identifier plusieurs comportements à risque, tels que visiter un nombre élevé de sites Web, certaines catégories de sites Web, un volume élevé d’activités en ligne, la participation à des réseaux pair à pair, et un volume élevé de téléchargements de fichiers exécutables d’Internet.

**Facteurs démographiques** Une étude cas-témoins a été réalisée afin d’évaluer l’effet de l’âge et du genre sur le risque d’attaque par logiciels malveillants. Notre étude, présentée au Chapitre 6, a été réalisée sur une période de deux mois durant lesquels nous avons suivi l’état de *santé* de plusieurs systèmes. La maladie, dans notre cas, était le fait d’avoir été victime d’au moins une attaque par logiciels malveillants. Les données concernant les attaques ont été collectées par Windows Defender sur plus de trois millions de systèmes durant deux mois. Ces informations ont par la suite été couplées avec les données de Microsoft Account afin d’obtenir le groupe d’âge et le genre associé au compte usager. Ce travail vient contribuer à la littérature existante en supportant que i) l’âge et le genre ont un impact indépendant sur le risque d’attaque par logiciels malveillants, et que ii) l’impact de l’âge et du genre varient en direction et importance en fonction du type de logiciel malveillant.

### 10.1.3 Évaluation de stratégie

*Évaluer l'efficacité réelle d'une intervention visant à prévenir et/ou réduire l'occurrence des attaques par logiciels malveillants.*

**Évaluation agrégée** Nous nous sommes inspirés du concept de santé des écosystèmes afin d'étudier la performance agrégée des solutions antivirus. Tel que présenté au Chapitre 7, nous avons considéré la santé d'un écosystème de logiciels antivirus comme étant sa performance agrégée à protéger les ordinateurs contre les attaques par logiciels malveillants. Nous avons à cet effet développé des indicateurs reliés à l'activité, la diversité, et la stabilité de l'écosystème. L'activité a été définie comme le pourcentage de systèmes qui sont activement protégés par une solution antivirus à jour. La diversité a été évaluée par la richesse spécifique, le degré de concentration, et la dominance. Finalement, la stabilité a permis de mesurer au sein des systèmes les changements au niveau des solutions antivirus et de leur état. L'ensemble de ces informations a été collecté par l'outil MSRT sur plus d'un milliard de systèmes Windows durant une période de quatre mois. Nos principales contributions issues de cette recherche sont i) le développement du concept de santé d'écosystème pour l'environnement des solutions antivirus, ii) la définition et la mesure d'indicateurs de santé et iii) l'évaluation de la performance agrégée de solutions antivirus.

**Évaluation comparative** Le test d'antivirus développé au Chapitre 8 a pris la forme d'une étude de cohorte. Dans notre cas, la population cible était composée d'utilisateurs de Windows 10, la maladie consistait à avoir été infecté par un logiciel malveillant, et nous avons considéré le fait d'être protégé par une solution antivirus comme étant notre facteur d'exposition. Au total, plus de 26 millions d'utilisateurs Windows 10 ont été étudiés pour une période de quatre mois. Les données concernant les infections par logiciels malveillants pour le groupe protégé ont été collectées par l'outil MSRT, et par Microsoft Windows Defender pour le groupe de comparaison. Ces informations ont par la suite été couplées avec Microsoft Account pour obtenir les facteurs au niveau de l'utilisateur (âge et genre). Quant aux facteurs au niveau de l'environnement, nous avons considéré la région géographique (par exemple, Amérique du Nord, Europe, Australie) ainsi que l'indice de développement humain. En comparant la fréquence d'infection associée à chaque groupe, nous avons été en mesure d'estimer l'efficacité réelle d'une solution antivirus à prévenir les infections par logiciels malveillants. Les principales contributions de ce travail sont i) le développement d'une méthodologie novatrice permettant d'évaluer l'efficacité réelle des solutions antivirus, et ii) la mise en évidence que la performance des solutions antivirus varie significativement en fonction de facteurs externes,

tels que le contexte socio-économique, le profil de l'utilisateur, et le type de logiciels malveillants.

## 10.2 Limitations et travaux futurs

Les principaux travaux de cette thèse —le développement d'un modèle de prévention, l'identification de déterminants, et l'évaluation d'une stratégie de prévention, ne sont que des premiers pas qui ouvrent la voie à plusieurs avenues de recherche. Notamment, les travaux suivants s'inscrivent dans la suite logique de la présente thèse :

### Identification des déterminants

- *Réaliser des études observationnelles basées sur différentes sources de données et périodes de temps afin de valider la généralisation de nos résultats.* Une des principales sources de limitations à l'égard de la validité externe de nos résultats résulte de nos sources de données. Premièrement, nos travaux de recherche portent sur l'écosystème du système d'exploitation Windows. Ainsi, nos résultats peuvent ne pas être représentatifs des autres systèmes d'exploitation, tels que Linux ou MacOS. En particulier, il est possible que les utilisateurs et les logiciels malveillants varient d'un système d'exploitation à l'autre. Deuxièmement, nos sources de données ont été collectées entre 2011 et 2016. Compte tenu de l'évolution rapide du problème —les attaques par logiciels malveillants— il est possible que nos résultats ne soient pas généralisables à d'autres périodes dans le temps. Troisièmement, nos indicateurs d'attaques par logiciels malveillants (détections ou infections) sont fonction des produits de sécurité utilisés (Trend Micro, MSRT, Microsoft Windows Defender) dans le cadre de nos travaux. Il est possible que les attaques observées aient été sous-estimées ou surestimées, que le comportement des usagers varie d'un produit à l'autre, ou encore que certains produits diffèrent quant à leur capacité et choix de détection. Somme toute, compte tenu de la popularité du système d'exploitation Windows, nous sommes d'avis que les résultats obtenus sont importants en soit, qu'ils soient représentatifs ou non du contexte des autres systèmes.
- *Étendre l'étude des déterminants reliés au risque d'attaques par logiciels malveillants afin de considérer un éventail plus large de facteurs.* Bien que nous ayons étudié plusieurs déterminants au niveau du système, de l'utilisateur, de l'environnement et des politiques, il est possible qu'il existe d'autres déterminants reliés aux attaques par logiciels malveillants qui n'ont pas été capturés par nos analyses. Il serait par conséquent intéressant d'envisager l'étude de facteurs additionnels, tels que la culture, la



communauté, ou l'investissement privé en sécurité des systèmes d'information.

- *Étudier la contribution relative des déterminants identifiés.* Nos travaux sur les déterminants portent principalement sur un niveau de facteurs à la fois. Une prochaine étape consiste à prendre en compte l'ensemble des facteurs identifiés. De telles études permettront de mesurer quelle est la contribution relative de chaque niveau (système, usager, environnement et politiques), et de prioriser les efforts de protection et de prévention.
- *Réaliser des études expérimentales ou quasi-expérimentales afin de confirmer la nature du lien causal entre les déterminants identifiés, et le risque d'attaques par logiciels malveillants.* Compte tenu de la nature observationnelle de nos études, les résultats obtenus sont potentiellement limités à l'identification de corrélats et non de causes. Par conséquent, la suite logique consiste à réaliser des études dites expérimentales ou quasi-expérimentales qui permettront de valider nos hypothèses étiologiques.

## Évaluation de stratégies

- *Réaliser des études utilisateurs afin de comprendre l'impact de l'utilisateur sur la performance des solutions antivirus.* Nous avons, dans le cadre de nos travaux, mis en évidence comment l'efficacité des solutions antivirus semble varier selon les caractéristiques et le comportement des utilisateurs. Bien que plusieurs hypothèses étiologiques aient été énoncées, l'identification des causes sous-jacentes permettrait d'améliorer les solutions existantes afin d'offrir une meilleure protection et ce, pour l'ensemble des usagers.
- *Évaluer l'efficacité réelle de différentes stratégies visant à prévenir les attaques par logiciels malveillants.* Nos travaux sur l'évaluation de stratégies ont porté sur une seule solution ; les produits antivirus. Lors de travaux subséquents, il serait pertinent d'étendre l'application à d'autres méthodes de prévention, telles que les mesures légales, la formation et l'éducation des usagers, ou encore les logiciels pare-feu.

## Modèle de prévention

- *Étendre l'application du modèle de prévention développé à l'implémentation et la promotion de stratégies de prévention des attaques par logiciels malveillants.* Notre application du modèle de prévention a principalement été axée sur les trois premières étapes, soit la définition du problème, l'identification des déterminants, et l'évaluation de stratégies. Dans cette optique, la prochaine étape consiste à appliquer notre modèle

de prévention au cas spécifique de la promotion des stratégies existantes de prévention des attaques par logiciels malveillants.

- *Développer des modèles de prévention applicables à d'autres types d'attaques informatiques.* La présente thèse est dédiée au cas spécifique des attaques par logiciels malveillants. Les résultats obtenus sont par conséquent limités en termes de généralisation à d'autres types de menaces informatiques. Une future avenue de recherche serait de s'inspirer de l'approche de la santé publique afin de développer des modèles de prévention qui sont adaptés aux spécificités des autres types de menaces.

### 10.3 Conclusion

Nous avons, dans le cadre de la présente thèse, développé et appliqué un modèle inspiré de la santé publique pour la prévention des attaques par logiciels malveillants. Plus particulièrement, nous nous sommes concentrés sur l'identification de déterminants reliés aux attaques par logiciels malveillants, et sur l'évaluation de l'efficacité réelle des solutions antivirus à prévenir l'occurrence des attaques par logiciels malveillants. Nos travaux ont notamment permis d'identifier plusieurs déterminants, et de mettre en lumière l'importance du contexte lorsque vient le temps de développer et d'évaluer des stratégies de prévention. Finalement, nous espérons que notre travail montre la valeur de s'inspirer des méthodes en santé publique et ce, non seulement pour la prévention des attaques par logiciels malveillants, mais pour d'autres fonctions au-delà de la prévention, ainsi que d'autres types d'attaques informatiques.

## RÉFÉRENCES

- Agence de la santé publique du Canada (2011). Qu'est-ce qui détermine la santé? <http://www.phac-aspc.gc.ca/ph-sp/determinants/index-fra.php>.
- Akinwande, Michael Olusegun and Dikko, Hussaini Garba and Samson, Agboola (2015). Variance inflation factor : As a condition for the inclusion of suppressor variable (s) in regression analysis. *Open Journal of Statistics*, 5(07), 754.
- Alam, Shahid and Sogukpinar, Ibrahim and Traore, Issa and Coady, Yvonne (2014). In-cloud malware analysis and detection : State of the art. *Proceedings of the 7th International Conference on Security of Information and Networks*. ACM, 473.
- Andrés, Antonio Rodríguez (2006). Software piracy and income inequality. *Applied Economics Letters*, 13(2), 101–105.
- Anthe, C and Chrzan, Patti (2015). Microsoft Security Intelligence Report January-June 2015. Rapport technique, Microsoft Corporation.
- Anti-Malware Testing Standards Organization (2008). Best practices for dynamic testing. <http://www.amtso.org/download/amtso-best-practices-for-dynamic-testing/?wpdmdl=1149>.
- Anti-Malware Testing Standards Organization Inc. (2016). Issues Involved in the Creation of Samples for Testing. Rapport technique, Anti-Malware Testing Standards Organization Inc.
- Asghari, Hadi and Ciere, Michael and Van Eeten, Michel JG (2015). Post-mortem of a zombie : conficker cleanup after six years. *Proceedings of the 24th USENIX Security Symposium*. 1–16.
- AV Comparatives (2013). File detection test of malicious software. Rapport technique, AV Comparatives.
- AV-TEST (2016). Security Report 2015/2016. [https://www.av-test.org/fileadmin/pdf/security\\_report/AV-TEST\\_Security\\_Report\\_2015-2016.pdf](https://www.av-test.org/fileadmin/pdf/security_report/AV-TEST_Security_Report_2015-2016.pdf).
- AV-TEST (2017). Security Report 2016/2017. [https://www.av-test.org/fileadmin/pdf/security\\_report/AV-TEST\\_Security\\_Report\\_2016-2017.pdf](https://www.av-test.org/fileadmin/pdf/security_report/AV-TEST_Security_Report_2016-2017.pdf).
- Berger, Wolfgang H and Parker, Frances L (1970). Diversity of planktonic foraminifera in deep-sea sediments. *Science*, 168(3937), 1345–1347.
- Bertollo, Pietro (1998). Assessing ecosystem health in governed landscapes : a framework for developing core indicators. *Ecosystem health*, 4(1), 33–51.

- Bieberstein, Andrea (2013). *An Investigation of Women's and Men's Perceptions and Meanings Associated with Food Risks*. Springer Science & Business Media.
- Bilge, Leyla and Han, Yufei and Dell'Amico, Matteo (2017). Riskteller : Predicting the risk of cyber incidents. *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*. ACM, 1299–1311.
- Blackbird, Joe and Pfeifer, Bill (2013a). The global impact of anti-malware protection state on infection rates. *Proceedings of the 23th Virus Bulletin International Conference*.
- Blackbird, J and Pfeifer, B (2013b). The global impact of anti-malware protection state on infection rates. *Proceedings of the 23th Virus Bulletin Conference*.
- Bossler, Adam M and Holt, Thomas J (2009). On-line activities, guardianship, and malware infection : An examination of routine activities theory. *International Journal of Cyber Criminology*, 3(1), 400.
- Burt, David and Nicholas, Paul and Sullivan, Kevin and Scoles, Travis (2014). The Cybersecurity Risk Paradox : Impact of social, economic, and technological factors on rates of malware. Rapport technique, Microsoft.
- Business Software Alliance (2012). 2011 BSA Global Software Piracy Study. Rapport technique, Business Software Alliance.
- Byrnes, James P and Miller, David C and Schafer, William D (1999). Gender differences in risk taking : A meta-analysis. *Psychological bulletin*, 125(3), 367.
- Canali, Davide and Bilge, Leyla and Balzarotti, Davide (2014). On the effectiveness of risk prediction based on users browsing behavior. *Proceedings of the 9th ACM symposium on Information, computer and communications security*. ACM, 171–182.
- Y. Carlinet and L. Mé and H. Débar and Y. Gourhant (2008). Analysis of computer infection risk factors based on customer network usage. *Second International Conference on Emerging Security Information, Systems and Technologies (SECURWARE'08)*. 317–325.
- Charney, Scott (2010). Collective defense : Applying public health models to the internet. *white paper (Redmond, Wash. : Microsoft Corporation, 2010)*, <http://www.microsoft.com/security/internethealth>.
- Chavez-Dreyfus, Gertrude (2018). About \$1.2 billion in cryptocurrency stolen since 2017 : cybercrime group. <https://www.reuters.com/article/us-crypto-currency-crime/about-1-2-billion-in-cryptocurrency-stolen-since-2017-cybercrime-group-idUSKCN1IP2LU>.
- Cherepanov, Anton (2017). Win32/Industroyer : A new threat for industrial control systems. Rapport technique, Eset.

- Choi, Kyung-shick (2008). Computer crime victimization and integrated theory : An empirical assessment. *International Journal of Cyber Criminology*, 2(1), 308.
- Cisco Systems (2017a). 2017 annual cybersecurity report. [https://www.cisco.com/c/dam/m/digital/1198689/Cisco\\_2017\\_ACR\\_PDF.pdf](https://www.cisco.com/c/dam/m/digital/1198689/Cisco_2017_ACR_PDF.pdf).
- Cisco Systems (2017b). Cisco Visual Networking Index : Forecast and Methodology, 2016-2021. <https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/complete-white-paper-c11-481360.pdf>.
- Conseil canadien des déterminants de la santé (2015). Examen des cadres relatifs aux déterminants de la santé. [http://ccsdh.ca/images/uploads/Examen\\_des\\_cadres.pdf](http://ccsdh.ca/images/uploads/Examen_des_cadres.pdf).
- Microsoft Corporation (2017). Microsoft Security Intelligence Report Volume 22 January-March 2017.
- Costanza, Robert and Norton, Bryan G and Haskell, Benjamin D (1992). *Ecosystem health : new goals for environmental management*. Island Press.
- Dahlgren, Göran and Whitehead, Margaret (1991). Policies and strategies to promote social equity in health. *Stockholm : Institute for future studies*.
- Dohmen, Thomas J. and Falk, Armin and Huffman, David and Sunde, Uwe and Schupp, Jürgen and Wagner, Gert G (2005). Individual risk attitudes : New evidence from a large, representative, experimentally-validated survey. *Social Science Research Network*.
- Dohmen, Thomas J. and Falk, Armin and Huffman, David and Sunde, Uwe and Schupp, Jürgen and Wagner, Gert G (2011). Individual risk attitudes : Measurement, determinants, and behavioral consequences. *Journal of the European Economic Association*, 9(3), 522–550.
- Dumitras, Tudor (2017). Field data available at symantec research labs : The worldwide intelligence network environment (wine).
- Dumitras, Tudor and Shou, Darren (2011). Toward a standard benchmark for computer security research : The worldwide intelligence network environment (wine). *Proceedings of the First Workshop on Building Analysis Datasets and Gathering Experience Returns for Security*. ACM, 89–96.
- Edwards, Benjamin and Hofmeyr, Steven and Forrest, Stephanie and van Eeten, Michel (2015). Analyzing and modeling longitudinal security data : Promise and pitfalls. *Proceedings of the 31st Annual Computer Security Applications Conference*. ACM, 391–400.
- Edwards, Simon PG (2013). Four Fs of anti-malware testing : A practical approach to testing endpoint security products. *Anti-malware Testing Research (WATeR), 2013 Workshop on*. IEEE, 1–9.

- Egelman, Serge and Peer, Eyal (2015). The myth of the average user : Improving privacy and security systems through individualization. *Proceedings of the 2015 New Security Paradigms Workshop*. ACM, 16–28.
- Emond, Catherine and Tang, Wei and Handy, Susan (2009). Explaining gender difference in bicycling behavior. *Transportation Research Record : Journal of the Transportation Research Board*, 2125, 16–25.
- Eset (2016). En Route with Sednit. Rapport technique, Eset.
- Eurostat (2011). Nearly one third of internet users in the EU27 caught a computer virus. [http://epp.eurostat.ec.europa.eu/cache/ITY\\_PUBLIC/4-07022011-AP/EN/4-07022011-AP-EN.PDF](http://epp.eurostat.ec.europa.eu/cache/ITY_PUBLIC/4-07022011-AP/EN/4-07022011-AP-EN.PDF).
- Feinberg, Robert M (1977). Risk aversion, risk, and the duration of unemployment. *The Review of Economics and Statistics*, 264–271.
- Figner, Bernd and Mackinlay, Rachael J and Wilkening, Friedrich and Weber, Elke U (2009). Affective and deliberative processes in risky choice : age differences in risk taking in the columbia card task. *Journal of Experimental Psychology : Learning, Memory, and Cognition*, 35(3), 709.
- Fischer, Justina AV and Rodriquez Andrés, Antonio (2005). Is software piracy a middle class crime ? investigating the inequality-piracy channel. *University of St. Gallen Economics Discussion Paper*, (2005-18).
- Forget, Alain and Komanduri, Saranga and Acquisti, Alessandro and Christin, Nicolas and Cranor, Lorrie Faith and Telang, Rahul (2014). Building the security behavior observatory : an infrastructure for long-term monitoring of client machines. *Proceedings of the 2014 Symposium and Bootcamp on the Science of Security*. ACM, 24.
- Forrest, Stephanie and Somayaji, Anil and Ackley, David H (1997). Building diverse computer systems. *The Sixth Workshop on Hot Topics in Operating Systems*. IEEE, 67–72.
- Fryer, Huw (2012). A public health approach to cybersecurity.
- Furnell, Steven (2010). Usability versus complexity—striking the balance in end-user security. *Network Security*, 2010(12), 13–17.
- Garg, Vaibhav and Koster, Thomas and Camp, Linda J (2013). Cross-country analysis of spambots. *EURASIP Journal on Information Security*, 2013(1), 3.
- Geer, Daniel and Bace, Rebecca and Gutmann, Peter and Metzger, Perry and Pfleeger, Charles P and Quarterman, John S and Schneier, Bruce (2003). Cyberinsecurity : The cost of monopoly. *How the dominance of Microsoft's products poses a risk to society*.

- Goel, Rajeev K and Nelson, Michael A (2009). Determinants of software piracy : economics, institutions, and technology. *The Journal of Technology Transfer*, 34(6), 637–658.
- Goel, Sharad and Hofman, Jake M and Siner, M Irmak (2012). Who does what on the web : A large-scale study of browsing behavior. *ICWSM*.
- S. Gordon and R. Ford (1996). Real world anti-virus product reviews and evaluations : the current state of affairs. *National Information Systems Security Conference*.
- Grimes, Galen A and Hough, Michelle G and Signorella, Margaret L (2007). Email end users and spam : relations of gender and age group to attitudes and actions. *Computers in Human Behavior*, 23(1), 318–332.
- Harley, David (2009). Making sense of anti-malware comparative testing. *information security technical report*, 14(1), 7–15.
- Harley, David and Lee, A. (2008). Who will test the testers. *Proceedings of the 18th Virus Bulletin International Conference*. 199–207.
- Harris, Christine R and Jenkins, Michael and Glaser, Dale (2006). Gender differences in risk assessment : why do women take fewer risks than men? *Judgment and Decision Making*, 1(1), 48.
- Hassan, Jahanzaib (2017). Anonymous hacks ISIS website; infecting users with malware. <https://www.hackread.com/anonymous-hacks-isis-site-with-malware/>.
- Hersch, Joni and Viscusi, W Kip (1990). Cigarette smoking, seatbelt use, and differences in wage-risk tradeoffs. *Journal of Human Resources*, 202–227.
- Hoffman, Chris (2016). Why Windows Has More Viruses than Mac and Linux. <https://www.howtogeek.com/141944/htg-explains-why-windows-has-the-most-viruses/>.
- Hogarth, Robin M and Portell, Mariona and Cuxart, Anna (2007). What risks do people perceive in everyday life? A perspective gained from the experience sampling method (esm). *Risk Analysis*, 27(6), 1427–1439.
- Hu, Jian and Zeng, Hua-Jun and Li, Hua and Niu, Cheng and Chen, Zheng (2007). Demographic prediction based on user's browsing behavior. *Proceedings of the 16th international conference on World Wide Web*. ACM, 151–160.
- Iansiti, Marco and Levien, Roy (2004a). Keystones and dominators : Framing operating and technology strategy in a business ecosystem. *Harvard Business School, Boston*.
- Iansiti, Marco and Levien, Roy (2004b). Strategy as ecology. *Harvard business review*, 82(3), 68–81.
- Iansiti, Marco and Richards, Gregory L (2006). Information technology ecosystem : Structure, health, and performance, the. *Antitrust Bull.*, 51, 77.

International Data Corporation (2013). Unlicensed Software and Cybersecurity Threats. [http://globalstudy.bsa.org/2013/Malware/study\\_malware\\_en.pdf](http://globalstudy.bsa.org/2013/Malware/study_malware_en.pdf).

International Secure Systems Lab (2013). Anubis Malware Analysis for Unknown Binaries. <https://anubis.iseclab.org/>.

International Telecommunication Union (2015). World Telecommunication/ICT Indicators database. <http://www.itu.int/en/ITU-D/Statistics/Pages/publications/wtid.aspx>.

Jagatic, Tom N and Johnson, Nathaniel A and Jakobsson, Markus and Menczer, Filippo (2007). Social phishing. *Communications of the ACM*, 50(10), 94–100.

Jansen, Slinger (2014). Measuring the health of open source software ecosystems : Beyond the scope of project health. *Information and Software Technology*, 56(11), 1508–1519.

Johnson, Joseph and Wilke, Andreas and Weber, Elke U (2004). Beyond a trait view of risk taking : A domain-specific scale measuring risk perceptions, expected benefits, and perceived-risk attitudes in german-speaking populations. *Polish Psychological Bulletin*, 35, 153–172.

Joiner, Richard and Gavin, Jeff and Brosnan, Mark and Cromby, John and Gregory, Helen and Guiller, Jane and Maras, Pam and Moon, Amy (2012). Gender, internet experience, internet identification, and internet anxiety : a ten-year followup. *Cyberpsychology, Behavior, and Social Networking*, 15(7), 370–372.

JPCERT Coordination Center (2014). The cyber green initiative : Improving health through measurement and mitigation. [https://www.jpccert.or.jp/research/GreenConcept-20141117\\_en.pdf](https://www.jpccert.or.jp/research/GreenConcept-20141117_en.pdf).

JPCERT Coordination Center (2015). Cyber green research paper. <http://static1.squarespace.com/static/54caa8ffe4b0184795b6296b/t/54de07fde4b05bfbd8825f63/1423837188601/Cyber+Green+Research+Paper++2015.pdf>.

Kalafut, Andrew and Acharya, Abhinav and Gupta, Minaxi (2006). A study of malware in peer-to-peer networks. *Proceedings of the 6th ACM SIGCOMM conference on Internet measurement*. ACM, 327–332.

Kaspersky Lab (2016). Kaspersky Lab Number of the Year 2016 : 323,00 pieces of malware detected daily. [https://usa.kaspersky.com/about/press-releases/2016\\_kaspersky-lab-number-of-the-year-2016-323000-pieces-of-malware-detected-daily](https://usa.kaspersky.com/about/press-releases/2016_kaspersky-lab-number-of-the-year-2016-323000-pieces-of-malware-detected-daily).

Kaspersky Lab (2017). Kaspersky Lab Number of the year : 360,00 Malicious Files Detected Daily in 2017. [https://usa.kaspersky.com/about/press-releases/2017\\_kaspersky-lab-number-of-the-year](https://usa.kaspersky.com/about/press-releases/2017_kaspersky-lab-number-of-the-year).



- Kaspersky Lab (2018). Adoption Rate and Popularity. <https://usa.kaspersky.com/resource-center/threats/malware-popularity>.
- Kindig, David and Stoddart, Greg (2003). What is population health? *American Journal of Public Health*, 93(3), 380–383.
- Kleiner, Aaron and Nicholas, Paul and Sullivan, Kevin (2013). Linking Cybersecurity Policy and Performance. Rapport technique, Microsoft Trustworthy Computing.
- P. Kosinar and J. Malcho and R. Marko and and D. Harley (2010). AV testing exposed. *20th Virus Bulletin International Conference*.
- Kranenbarg, Marleen Weulen and Holt, Thomas J and van Gelder, Jean-Louis (2017). Offending and victimization in the digital age : Comparing correlates of cybercrime and traditional offending-only, victimization-only and the victimization-offending overlap. *Deviant Behavior*, 1–16.
- Kumar, Ajay and Ojha, Nishikant and Srivastava, Nishit Kumar (2017). Factors affecting malware attacks : An empirical analysis. *Purushartha : A Journal of Management Ethics and Spirituality*, 10(2).
- Kumaraguru, Ponnurangam and Cranshaw, Justin and Acquisti, Alessandro and Cranor, Lorrie and Hong, Jason and Blair, Mary Ann and Pham, Theodore (2009). School of phish : a real-world evaluation of anti-phishing training. *Proceedings of the 5th Symposium on Usable Privacy and Security*. ACM, 3.
- Lalonde Lévesque, Fanny and Chiasson, Sonia and Somayaji, Anil and Fernandez, José M. (2018). Technological and human factors of malware attacks : A computer security clinical trial approach. *ACM Transactions on Privacy and Security (TOPS)*, 21(4), 18.
- Lalonde Lévesque, Fanny and Davis, C.R. and Fernandez, J.M. and Chiasson, S. and Somayaji, A. (2012a). Methodology for a field study of anti-malware software. *Workshop on Usable Security (USEC)*. LNCS, 80–85.
- Lalonde Lévesque, Fanny and Davis, C.R. and Fernandez, J.M. and Somayaji, A. (2012b). Evaluating antivirus products with field studies. *22th Virus Bulletin International Conference*. 87–94.
- Lalonde Lévesque, Fanny and Fernandez, José M (2014). Computer security clinical trials : Lessons learned from a 4-month pilot study. *USENIX Workshop on Cyber Security Experimentation and Test (CSET)*. ACM.
- Lalonde Lévesque, Fanny and Fernandez, José M and Batchelder, Dennis (2017). Age and gender as independent risk factors for malware victimisation. *Proceedings of the 31th International British Human Computer Interaction Conference*.

- Lalonde Lévesque, Fanny and Fernandez, José M. and Somayaji, Anil (2014). Risk prediction of malware victimization based on user behavior. *Malicious and Unwanted Software : The Americas (MALWARE), 2014 9th International Conference on*. IEEE, 128–134.
- Lalonde Lévesque, Fanny and Fernandez, José M. and Somayaji, Anil and Batchelder, Dennis (2016). National-level risk assessment : A multi-country study of malware infections. *Proceedings of the 15th Workshop on The Economics of Information Security*.
- Lalonde Lévesque, Fanny and Fernandez, José M. and Batchelder, Dennis and Young, Glau-  
cia (2016a). Are they real? real-life comparative tests of anti-virus products. *Proceedings of the 26th International Virus Bulletin Conference*.
- Lalonde Lévesque, Fanny and Fernandez, José M. and Batchelder, Dennis and Young, Glau-  
cia (2016b). Are they real? real-life comparative tests of anti-virus products. *26th Virus Bulletin International Conference*. 25–33.
- Lalonde Lévesque, Fanny and Nsiempba, Jude and Fernandez, José M. and Chiasson, Sonia and Somayaji, Anil (2013). A clinical study of risk factors related to malware infections. *Proceedings of the 2013 ACM SIGSAC Conference on Computer & Communications Security*. ACM, 97–108.
- Lalonde Lévesque, Fanny and Somayaji, Anil and Batchelder, Dennis and Fernandez, Jose M. (2015). Measuring the health of antivirus ecosystems. *Malicious and Unwanted Software : The Americas (MALWARE), 2015 10th International Conference on*. IEEE.
- Lee, Martin (2012). Who’s next? identifying risks factors for subjects of targeted attacks. *Proceedings of the 22th Virus Bulletin International Conference*. 301–306.
- Leukfeldt, Eric Rutger (2015). Comparing victims of phishing and malware attacks : Unraveling risk factors and possibilities for situational crime prevention. *arXiv preprint arXiv :1506.00769*.
- Leukfeldt, Eric Rutger and Yar, Majid (2016). Applying routine activity theory to cyber-crime : A theoretical and empirical analysis. *Deviant Behavior*, 37(3), 263–280.
- Maier, Gregor and Feldmann, Anja and Paxson, Vern and Sommer, Robin and Vallentin, Matthias (2011). An assessment of overt malicious activity manifest in residential networks. *Detection of Intrusions and Malware, and Vulnerability Assessment*, Springer. 144–163.
- Malik, Khalid and Jespersen (2013). Human Development Report 2013. Rapport technique, United Nations Development Programme.
- Manes, Casper (2015). 2015’s MVPs – The most vulnerable players. <https://techtalk.gfi.com/2015s-mvps-the-most-vulnerable-players/>.
- Manikas, Konstantinos and Hansen, Klaus Marius (2013). Reviewing the health of software ecosystems-a conceptual framework proposal. *IWSECO@ ICSOB*. 33–44.

- Marx, Andreas (2000). A guideline to anti-malware-software testing. *European Institute for Computer Anti-Virus Research (EICAR)*, 218–253.
- Mather, Mara (2006). A review of decision-making processes : Weighing the risks and benefits of aging. *When I ?m*, 64(145), 145–173.
- Mavrommatis, Niels Provos Panayiotis and Monroe, Moheeb Abu Rajab Fabian (2008). All your iframes point to us. *USENIX Security Symposium*. 1–16.
- McAfee Labs (2012). McAfee threats report : First quarter 2012. <http://www.mcafee.com/us/resources/reports/rp-quarterly-threat-q1-2012.pdf>.
- Menting, Michela (2014). Global Cybersecurity Index. Rapport technique, ABI Research.
- Mezzour, Ghita and Carley, Kathleen M and Carley, L Richard (2015). An empirical study of global malware encounters. *Proceedings of the 2015 Symposium and Bootcamp on the Science of Security*. ACM, 8.
- Microsoft (2017). Microsoft Privacy Statement. <https://privacy.microsoft.com/en-gb/privacystatement>.
- Microsoft Corporation (2013a). Microsoft Security Intelligence Report Volume 15 January through June, 2013. Rapport technique, Microsoft Corporation.
- Microsoft Corporation (2013b). Microsoft Security Intelligence Report Volume 16 July through December, 2013. Rapport technique, Microsoft Corporation.
- Microsoft Corporation (2014a). Microsoft Security Intelligence Report Volume 14 July through December, 2012. Rapport technique, Microsoft Corporation.
- Microsoft Corporation (2014b). Microsoft Security Intelligence Report Volume 17. Rapport technique.
- Microsoft Corporation (2014c). Microsoft Security Intelligence Report Volume 18. Rapport technique.
- Microsoft Corporation (2014d). Microsoft Security Intelligence Report Volume 18 July through December, 2014. Rapport technique, Microsoft Corporation.
- Microsoft Corporation (2015). Microsoft Security Intelligence Report Volume 20 July through December, 2015. Rapport technique, Microsoft Corporation.
- Microsoft Corporation (2016). Malware Families Cleaned by the Malicious Software Removal Tool. <http://www.microsoft.com/security/pc-security/malware-families.aspx>.
- Microsoft Secure (2017). Ransomware : A declining nuisance or an evolving menace? <https://cloudblogs.microsoft.com/microsoftsecure/2017/02/14/ransomware-2016-threat-landscape-review/>.

G. R. Milne and L. I. Labrecque and C. Cromer (2009). Toward an understanding of the online consumer's risky behavior and protection practices. *Journal of Consumer Affairs*, 43, 449–473.

Ministry of Health Services Province of British Columbia (2005). A framework for core functions in public health. [http://www.health.gov.bc.ca/library/publications/year/2005/core\\_functions.pdf](http://www.health.gov.bc.ca/library/publications/year/2005/core_functions.pdf).

Moar, James (2017). The Future of Cybercrime and Security : Enterprise Threats and Mitigation 2017-2022. <https://www.juniperresearch.com/researchstore/innovation-disruption/cybercrime-security/enterprise-threats-mitigation>.

Morgan, Steve (2017). 2017 Cybercrime Report. Rapport technique, Cybersecurity Ventures.

Mulligan, Deirdre K and Schneider, Fred B (2011). Doctrine for cybersecurity. *Daedalus*, 140(4), 70–92.

Muttik, Igor and Vignoles, James (2008). Rebuilding anti-malware testing for the future. *Virus Bulletin Conference*.

National Cyber Security Alliance (2018). National Cybersecurity Awareness Month. <https://staysafeonline.org/ncsam/>.

Ngo, Fawn T. and Paternoster, Raymond (2011). Cybercrime victimization : An examination of individual and situational level factors. *International Journal of Cyber Criminology*, 5(1), 773–793.

No More Ransom Project (2018). No More Ransom! <https://www.nomoreransom.org>.

O'Brien, Dick (2016). Survey Report : Understanding the Depth of the Global Ransomware Problem. <https://go.malwarebytes.com/OstermanRansomwareSurvey.html>.

O'Brien, Dick (2017). Internet Security Threat Report : Ransomware 2017. <https://www.symantec.com/content/dam/symantec/docs/security-center/white-papers/istr-ransomware-2017-en.pdf>.

Ogala, Emmanuel (2013). Gay Activist Hacks Nigerian Government's Web-site Over Country's Anti-Gay Law. <https://www.premiumtimesng.com/news/140270-gay-activist-hacks-nigerian-governments-website-over-countrys-anti-gay-law.html>

Oliveira, Daniela and Rocha, Harold and Yang, Huizi and Ellis, Donovan and Dommaraju, Sandeep and Muradoglu, Melis and Weir, Devon and Soliman, Adam and Lin, Tian and Ebner, Natalie (2017). Dissecting spear phishing emails for older vs young adults : On the interplay of weapons of influence and life domains in predicting susceptibility to phishing.

*Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, 6412–6424.

Onarlioglu, Kaan and Yilmaz, Utku Ozan and Kirda, Engin and Balzarotti, Davide (2012). Insights into user behavior in dealing with internet attacks. *NDSS*.

OPSWAT (2014). Antivirus and Threat Report : January 2014. <https://www.opswat.com/resources/reports/antivirus-january-2014>.

Ovelgönne, Michael and Dumitras, Tudor and Prakash, B Aditya and Subrahmanian, VS and Wang, Benjamin (2017). Understanding the relationship between human behavior and susceptibility to cyber attacks : A data-driven approach. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 8(4), 51.

Panda Labs (2012). PandaLabs Q1 Report : Four Out Of Five New Malware Samples Are Trojans.

Panda Security Labs (2011). Panda Labs annual report 2011 summary. <http://press.pandasecurity.com/wp-content/uploads/2012/01/Annual-Report-PandaLabs-2011.pdf>.

Panda Security Labs (2012). Panda Labs quarterly report April - June 2012. <https://www.pandasecurity.com/mediacenter/src/uploads/2012/08/Quarterly-Report-PandaLabs-April-June-2012.pdf>.

PandaLabs (2017). PandaLabs Annual Report 2017 . [https://www.pandasecurity.com/mediacenter/src/uploads/2017/11/PandaLabs\\_Annual\\_Report\\_2017.pdf](https://www.pandasecurity.com/mediacenter/src/uploads/2017/11/PandaLabs_Annual_Report_2017.pdf).

Parker, R David and Farkas, Csilla (2011). Modeling estimated risk for cyber attacks : Merging public health and cyber security. *Information Assurance and Security Letters*, 2, 32–6.

PC Security Labs (2013). Security solution review on Windows 8 platform. Rapport technique, PC Security Labs.

Png, Ivan PL and Wang, Chen-Yu and Wang, Qiu-Hong (2008). The deterrent and displacement effects of information security enforcement : International evidence. *Journal of Management Information Systems*, 25(2), 125–144.

Privacy Rights ClearingHouse (2018). Data Breaches. <https://www.privacyrights.org/data-breaches>.

Proofpoint (2016). Q4 2016 and Year in Review : Threat Summary. [https://www.proofpoint.com/sites/default/files/q4\\_threat-summary-final-cm-16217.pdf](https://www.proofpoint.com/sites/default/files/q4_threat-summary-final-cm-16217.pdf).

Provos, Niels and McNamee, Dean and Mavrommatis, Panayiotis and Wang, Ke and Modadugu, Nagendra and others (2007). The ghost in the browser : Analysis of web-based malware. *HotBots*, 7, 4–4.

- Rains, Tim (2013). The threat landscape in china : A paradox. <https://blogs.microsoft.com/cybertrust/2013/03/11/the-threat-landscape-in-china-a-paradox/>.
- Reyns, Bradford W (2013). Online routines and identity theft victimization further expanding routine activity theory beyond direct-contact offenses. *Journal of Research in Crime and Delinquency*, 50(2), 216–238.
- Rice, Mason and Butts, Jonathan and Miller, Robert and Shenoi, Sujeet (2010). Applying public health strategies to the protection of cyberspace. *International Journal of Critical Infrastructure Protection*, 3(3), 118–127.
- Rolison, Jonathan J and Hanoch, Yaniv and Wood, Stacey and Liu, Pi-Ju (2013). Risk-taking differences across the adult life span : a question of age and domain. *The Journals of Gerontology Series B : Psychological Sciences and Social Sciences*, gbt081.
- Rowe, Brent and Halpern, Michael and Lentz, Tony (2012a). Is a public health framework the cure for cyber security? *CrossTalk*, 25(6), 30–38.
- Rowe, Brent and Halpern, Michael and Lentz, Tony and Wood, Dallas (2012b). Understanding cyber security risk preferences : A case study analysis inspired by public health research.
- Rowe, Jeff and Levitt, Karl and Hogarth, Mike (2013). Towards the realization of a public health model for shared secure cyber-space. *Proceedings of the 2013 New Security Paradigms Workshop*. ACM.
- Saeed, Imtithal A and Selamat, Ali and Abuagoub, Ali MA (2013). A survey on malware and malware detection systems. *International Journal of Computer Applications*, 67(16).
- Samuelson, Paul A (1954). The pure theory of public expenditure. *The review of economics and statistics*, 387–389.
- Savage, Kevin and Coogan, Peter and Lau, Hon (2015). The evolution of ransomware. Rapport technique, Symantec.
- Sedenberg, Elaine M and Mulligan, Deirdre K (2015). Public health as a model for cybersecurity information sharing. *Berkeley Tech. LJ*, 30, 1687.
- SERENE-RISC (2018). SERENE-RISC : Réseau intégré sur la cybersécurité. <https://www.serene-risc.ca/>.
- S. Sheng and M. Holbrook and P. Kumaraguru and L. F. Cranor and J. Downs (2010). Who falls for phish? A demographic analysis of phishing susceptibility and effectiveness of interventions. *ACM Conference on Human Factors in Computing Systems (CHI)*. 373–382.
- Simpson, Edward H (1949). Measurement of diversity. *nature*.

- Slepogin, Nikita (2017). Dridex : A History of Evolution. <https://securelist.com/dridex-a-history-of-evolution/78531/>.
- A. Somayaji and Y. Li and H. Inoue and J.M. Fernandez and R. Ford (2009). Evaluating security products with clinical trials. *USENIX Workshop on Cyber Security Experimentation and Test (CSET)*.
- Sood, Aditya K and Enbody, Richard J (2011). Malvertising—exploiting web advertising. *Computer Fraud & Security*, 2011(4), 11–16.
- Statista (2015). Security software - statistics and Facts. <https://www.statista.com/topics/2208/security-software/>.
- Subrahmanian, VS and Ovelgonne, Michael and Dumitras, Tudor and Prakash, Aditya (2016). The global cyber-vulnerability report.
- Sullivan, Kevin and Bahl, Sanjay and Boyer, Chris (2012). The internet health model for cybersecurity.
- SurfRight (2009). 32% of computers still infected, despite presence of antivirus program. <http://www.surfright.nl/en/home/press/32-percent-infected-despite-antivirus>.
- Symantec Corporation (2012). Internet Security Threat Report 2011 Trends. [http://www.symantec.com/content/en/us/enterprise/other\\_resources/b-istr\\_main\\_report\\_2011\\_21239364.en-us.pdf](http://www.symantec.com/content/en/us/enterprise/other_resources/b-istr_main_report_2011_21239364.en-us.pdf).
- Teo, Thompson SH (2001). Demographic and motivation variables associated with internet usage activities. *Internet Research*, 11(2), 125–137.
- The WildList Organization International (2017). The WildList. <https://www.wildlist.org/>.
- The World Bank (2017). World development indicators. <http://data.worldbank.org/products/wdi>.
- Thonnard, Olivier and Bilge, Leyla and Kashyap, Anand and Lee, Martin (2014). Are you at risk? profiling organizations and individuals subject to targeted attacks.
- Thonnard, Olivier and Bilge, Leyla and Kashyap, Anand and Lee, Martin (2015). Are you at risk? profiling organizations and individuals subject to targeted attacks. *International Conference on Financial Cryptography and Data Security*. Springer, 13–31.
- Trend Micro (2012). Website classification. <http://solutionfile.trendmicro.com/solutionfile/Consumer/new-web-classification.html>.

Trend Micro (2017). Down but Not Out : A Look Into Recent Exploit Kit Activities. <https://blog.trendmicro.com/trendlabs-security-intelligence/a-look-into-recent-exploit-kit-activities/>.

United Nations Development Programme (2015a). Human Development Report 2015. Rapport technique, United Nations Development Programme.

United Nations Development Programme (2015b). Human Development Reports. <http://hdr.undp.org/en/data>.

Unucheck, Roman and Sinitsyn, Fedor and Parinov, Denis and Stolyarov, Vladislav (2017). IT threat evolution Q1 207. Statistics.

Van Eeten, Michel and Bauer, Johannes M and Asghari, Hadi and Tabatabaie, Shirin and Rand, Dave (2010). The role of internet service providers in botnet mitigation : An empirical analysis based on spam data. Rapport technique, OECD Publishing.

Van Lingen, Sonny and Palomba, Adrien and Lucassen, Garm (2013). On the software ecosystem health of open source content management systems. *5th International Workshop on Software Ecosystems (IWSECO 2013)*. 38.

Vasek, Marie and Moore, Tyler (2014). Identifying risk factors for webserver compromise. *International Conference on Financial Cryptography and Data Security*. Springer, 326–345.

Vasek, Marie and Wadleigh, John and Moore, Tyler (2016). Hacking is not random : a case-control study of webserver-compromise risk. *IEEE Transactions on Dependable and Secure Computing*, 13(2), 206–219.

Virus Total (2013). Virus Total. <https://www.virustotal.com>.

J. Vrabec and D. Harley (2010). Real performance? *EICAR Annual Conference*.

Wang, Mei and Keller, Carmen and Siegrist, Michael (2011). The less you know, the more you are afraid of? a survey on risk perceptions of investment products. *Journal of Behavioral Finance*, 12(1), 9–19.

Weber, Elke U and Blais, Ann-Renee and Betz, Nancy E (2002). A domain-specific risk-attitude scale : Measuring risk perceptions and risk behaviors. *Journal of behavioral decision making*, 15(4), 263–290.

Wilcox, Bruce A (2001). Ecosystem health in practice : emerging areas of application in environment and human health. *Ecosystem Health*, 7(4), 317–325.

Wilsem, Johan van (2013). Hacking and harassment—do they have something in common? comparing risk factors for online victimization. *Journal of Contemporary Criminal Justice*, 29(4), 437–453.

Winslow, C. (1920). The untilled fields of public health. *Science*, 23–33.



- Wnuk, Krzysztof and Manikas, Konstantinos and Runeson, Per and Lantz, Matilda and Weijden, Oskar and Munir, Hussan (2014). Evaluating the governance model of hardware-dependent software ecosystems—a case study of the axis ecosystem. *Software Business. Towards Continuous Value Delivery*, Springer. 212–226.
- World Health Organization (2015). Tobacco use Data by country. <http://apps.who.int/gho/data/node.main.65>.
- Wueest, Candid (2015). Financial threats 2015. <https://www.symantec.com/content/dam/symantec/docs/security-center/white-papers/financial-threats-15-en.pdf>.
- Wynn Jr, Donald and Boudreau, MC and Watson, R (2007). Assessing the health of an open source ecosystem. *Emerging Free and Open Source Software Practices*, 238–258.
- Xing, Xinyu and Meng, Wei and Lee, Byoungyoung and Weinsberg, Udi and Sheth, Anmol and Perdisci, Roberto and Lee, Wenke (2015). Understanding malvertising through ad-injecting browser extensions. *Proceedings of the 24th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 1286–1295.
- Yen, Ting-Fang and Heorhiadi, Victor and Oprea, Alina and Reiter, Michael K and Juels, Ari (2014). An epidemiological study of malware encounters in a large enterprise. *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*. ACM, 1117–1130.
- Zaikin, Roman and Barda, Dikla (2016). ImageGate : Check Point uncovers a new method for distributing malware through images. <https://blog.checkpoint.com/2016/11/24/imagegate-check-point-uncovers-new-method-distributing-malware-images/>.
- Zainodin, HJ and Khuneswari, G and Noraini, A and Haider, FAA (2015). Selected model systematic sequence via variance inflationary factor. *International Journal of Applied Physics and Mathematics*, 5(2), 105.
- Zelonis, Kim (2004). *Avoiding the cyber pandemic : A public health approach to preventing malware propagation*. Thèse de doctorat, Carnegie Mellon University.
- Zwienenberg, Righard and Ford, Richard and Wegele, Thomas (2013). The Real Time Threat List. *Proceedings of the 23th Virus Bulletin International Conference*.

## ANNEXE A    DESCRIPTION OF COUNTRY-LEVEL FACTORS

<b>Factor</b>	<b>Definition</b>
GDP per capita	GDP converted from domestic currencies to U.S. dollars using single year official exchange rates.
GDP per capita by purchasing power parity	GDP per capita based on purchasing power parity.
Mean years of schooling	Average number of years of education received by people ages 25 and older, converted from education attainment levels using official durations of each level.
%Households with computer	Percentage of households with computer.
%Households with Internet	Percentage of households with Internet.
Fixed (wired) Internet subscriptions (per 100 inhabitants)	Number of active fixed (wired) Internet subscriptions at speed less than 256 kbits/s and the total fixed (wired) broadband subscriptions.
Fixed (wired) broadband subscriptions (per 100 inhabitants)	Number of fixed (wired) broadband subscriptions with access over wireline networks. Wireless broadband is not included.
Fixed (wired) broadband speed	Refers to the advertised maximum theoretical download speed, and not speeds guaranteed to users associated with a fixed (wired) broadband Internet monthly subscriptions. It does not refer to the actual speed delivered.
Fixed broadband subscriptions between 256 kbits/s and less than 2 Mbits/s	Percentage of Internet broadband subscriptions with advertised downstream speed equal to 256 kbits/s and less than 2 Mbits/s.
Fixed broadband subscriptions between 2 Mbits/s and less than 10 Mbits/s	Percentage of Internet broadband subscriptions with advertised downstream speed equal to 2 Mbits/s and less than 10 Mbits/s.
Fixed broadband subscriptions above 10 Mbits/s	Percentage of Internet broadband subscriptions with advertised downstream speed equal to, or greater than 10 Mbits/s.
International Internet bandwidth	Total used capacity of international Internet bandwidth, in megabits per second. Measures the sum of used capacity of all Internet exchanges offering international bandwidth.
Secure Internet servers (per 1 million people)	Number of secure Internet servers using encryption technology in Internet transactions.
%Protected	Refers to the percentage of users that have at least one anti-malware product actively running with up-to-date signatures.
Global cybersecurity index	Index of level of cybersecurity development in terms of legal measures, technical measures, organizational measures, capacity building and cooperation.

## ANNEXE B RESIDUAL ANALYSIS

We first plotted the raw residuals versus the predicted values to examine if the raw residuals have a constant variance, and a mean of 0. As depicted in Figure B.1(a), the fitted line plot shows that the mean is 0 and that the assumption of equal variance does not seem to be violated. The plot of the expected normal value versus the raw residual was also examined to see if the residuals follow a normal distribution. Visual inspection of the plot (see Figure B.1(b)) suggests that the residuals follow a straight line, meaning that a linear regression model is adequate.

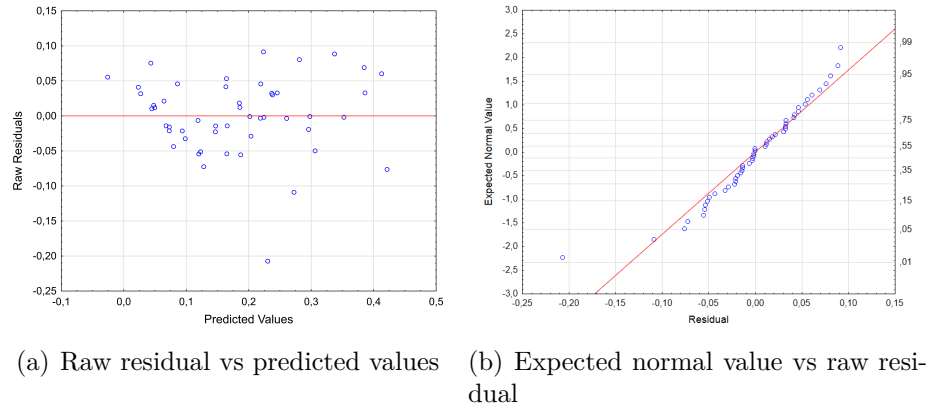


Figure B.1 Residual analysis

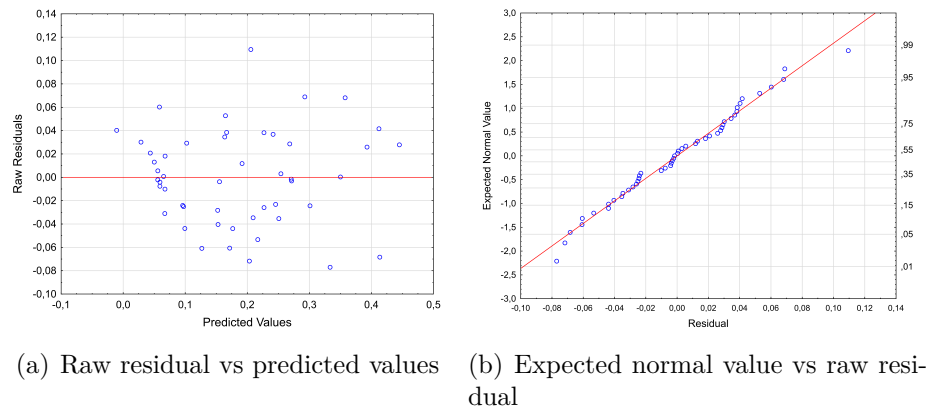


Figure B.2 Residual analysis without China

The graphical analysis also suggested that one observation (China) may be an outlier. We computed for each country the associated standardized residual (also known as the stu-

dentized residual) to identify potential outliers. In general, an absolute value larger than 3 indicates that the observation is an outlier. Results of the analysis (3.46) confirmed that China is an outlier according to our regression model. We then performed another residual analysis excluding China (see Figure B.2(a) and Figure B.2(b)), which also suggested that a linear regression model is appropriate for the data.

## ANNEXE C    MULTIPLE GENERAL LINEAR REGRESSION RESULTS

Table C.1 Global multiple general linear regression results (N=50 countries)

Factor	Developed (N=25)				Newly industrialized (N=22)				Developing (N=22)			
	$\beta$	Std.	t-value	p-value	$\beta$	Std.	t-value	p-value	$\beta$	Std.	t-value	p-value
	Error				Error				Error			
GDP-log	0.47	0.15	3.12	7.01e-03***	-0.91	0.23	-3.95	1.92e-03***	0.29	0.20	1.46	0.16
MYS	-0.08	0.11	-0.72	0.48	-0.71	0.17	-3.99	1.79e-03***	0.03	0.19	0.17	0.86
FBS-log	-0.24	0.08	-2.87	0.01**	0.12	0.16	0.74	0.47	0.16	0.21	0.75	0.46
%FB(256-2)-log	-0.19	0.11	-1.66	0.12	0.29	0.22	1.34	0.20	-	-	-	-
%FB(2-10)-log	-0.28	0.17	-1.63	0.22	0.67	0.21	3.25	6.95e-03***	-	-	-	-
%FB(10+)-log	-0.03	0.13	-1.02	-0.26	0.69	0.25	2.75	1.75e-02**	-	-	-	-
IIB-log	-0.31	0.11	-2.62	0.01**	-0.74	0.17	-4.45	7.87e-04***	0.28	0.24	1.17	0.26
%P	-0.80	0.17	-4.45	4.59e-04***	-0.51	0.14	-3.62	3.51e-03***	-0.72	0.021	-3.46	3.52e-03***
GCI	-0.21	0.09	-2.35	0.03*	0.30	0.17	1.83	0.09	-0.51	0.21	-2.40	0.03*
	R <sup>2</sup> adjusted				R <sup>2</sup> adjusted				R <sup>2</sup> adjusted			
	0.89				0.79				0.41			
	F-statistic				F-statistic				F-statistic			
	22.88				9.73				3.39			
	Df				Df				Df			
	9				9				6			
	Df (residuals)				Df (residuals)				Df (residuals)			
	15				21				15			
	p-value				p-value				p-value			
	3.41e-07				2.82e-04				0.02			

\*Statistically significant at 0.05 level; \*\*Statistically significant at 0.01 level; \*\*\*Statistically significant at 0.001 level.

## ANNEXE D    WINDOWS VERSIONS STATISTICS

Table D.1 Distribution of Windows versions by socio-economic status

Version	Developed		Newly industrialized		Developing	
	Mean	SD	Mean	SD	Mean	SD
Windows XP	0.18	0.06	0.14	0.08	0.14	0.06
Windows Vista	0.08	0.03	0.03	0.02	0.03	0.01
Windows 7	0.54	0.04	0.57	0.08	0.53	0.07
Windows 8	0.10	0.04	0.17	0.07	0.23	0.05
Windows 8.1	0.09	0.06	0.08	0.03	0.07	0.02

## ANNEXE E    DEFINITIONS BY MALWARE TYPE

**Adware** : Software that shows you extra promotions that you cannot control as you use your PC.

**Bot** : Small, hidden programs that are often controlled by a malicious hacker. Bots can be installed on your PC without you knowing.

**Cracks** : A type of tool that can be used to activate an unregistered copy of a software.

**Exploit** : A piece of code that uses software vulnerabilities to access information on your PC or install malware.

**Hack** : A type of tool that can be used to allow and maintain unauthorized access to your PC.

**Infostealer** : A type of malware that is used to steal your personal information, such as user names and passwords.

**Ransomware** : A type of malware that can stop you from using your PC, or encrypt your files so you cannot use them. You may be warned that you need to pay money, complete surveys, or perform other actions before you can use your PC again.

**Rogue** : Software that pretends to be an antivirus program but doesn't actually provide any security. This type of software usually gives you a lot of alerts about threats on your PC that don't exist. It also tries to convince you to pay for its services.

**Rootkit** : A program that is designed to hide itself and other malware from detection while it makes changes to your PC.

**Virus** : Type of malware that spread on their own by attaching their code to other programs, or copying themselves across systems and networks.

## ANNEXE F    ODDS RATIOS BY MALWARE TYPE

Table F.1 Odds ratios for gender by malware type

Malware	OR (95% CI)	$p$ – value
Adware	0.98 (0.97-0.99)	2.17e-10
Virus	1.66 (1.63-1.68)	< 1.00e-16
Cracks	2.01 (1.97-2.04)	< 1.00e-16
Hack	3.13 (2.88-3.40)	< 1.00e-16
Exploit	1.73 (1.65-1.82)	< 1.00e-16
Rogue	0.98 (0.94-1.02)	4.11e-01
Infostealer	1.97 (1.89-2.04)	< 1.00e-16
Ransomware	1.32 (1.24-1.40)	< 1.00e-16
Bot	2.09 (1.73-2.5)	< 1.00e-16
Rootkit	2.25 (1.92-2.64)	< 1.00e-16



Table F.2 Odds ratios for age by malware type

Malware	Age	OR(95% CI)	$p$ – value
Adware	0-17	1.75 (1.72-1.77)	< 1.00e-16
	18-24	1.36 (1.35-1.38)	< 1.00e-16
	25-34	1.01 (0.99-1.02)	0.00e-01
	35-49	0.99 (0.98-1.01)	< 1.00e-16
Virus	0-17	4.38 (4.21-4.56)	< 1.00e-16
	18-24	7.14 (6.93-7.37)	< 1.00e-16
	25-34	3.91 (3.79-4.03)	< 1.00e-16
	35-49	2.37 (2.29-2.46)	< 1.00e-16
Cracks	0-17	2.38 (2.31-2.45)	3.82e-11
	18-24	3.78 (3.70-3.86)	< 1.00e-16
	25-34	3.16 (3.09-3.22)	< 1.00e-16
	35-49	1.95 (1.91-1.99)	< 1.00e-16
Hack	0-17	4.40 (3.94-4.91)	< 1.00e-16
	18-24	3.18 (2.90-3.49)	< 1.00e-16
	25-34	2.45 (2.23-2.69)	8.15e-03
	35-49	1.88 (1.71-2.08)	1.93e-14
Exploit	0-17	0.76 (0.69-0.84)	< 1.00e-16
	18-24	1.10 (1.04-1.16)	6.49e-03
	25-34	1.27 (1.21-1.35)	< 1.00e-16
	35-49	1.16 (1.10-1.22)	1.37e-08
Rogue	0-17	0.51 (0.47-0.55)	2.56e-08
	18-24	0.47 (0.44-0.48)	< 1.00e-16
	25-34	0.54 (0.51-0.56)	1.71e-11
	35-49	0.64 (0.61-0.67)	4.02e-04
Infostealer	0-17	4.21 (3.94-4.49)	< 1.00e-16
	18-24	4.57 (4.33-4.81)	< 1.00e-16
	25-34	3.26 (3.09-3.44)	< 1.00e-16
	35-49	1.99 (1.88-2.10)	< 1.00e-16
Ransomware	0-17	0.68 (0.60-0.77)	1.70e-03
	18-24	0.65 (0.61-0.71)	4.60e-12
	25-34	0.76 (0.71-0.82)	9.02e-02
	35-49	0.93 (0.87-0.99)	2.15e-10
Bot	0-17	1.20 (1.05-1.38)	1.07e-01
	18-24	1.69 (1.55-1.84)	< 1.00e-16
	25-34	1.46 (1.34-1.59)	1.90e-05
	35-49	1.27 (1.17-1.39)	3.93e-01

## ANNEXE G    MULTIPLE LOGISTIC REGRESSION BY MALWARE TYPE

We present for each type of malware the results of the logistic regression. The odds ratio (OR) and its associated confidence interval (CI) at 95% were computed and are shown. We used the  $p$ -value as an indicator of whether the difference in exposure between the cases and the controls is statistically significant : \* indicates that the effect is statistically significant at 0.05 level ; \*\* at 0.01 level ; and \*\*\* at 0.001 level.

Table G.1 Multiple logistic regression for adware

Factor	Description	OR (95% CI)
Gender	Male	0.94 (0.93-0.95)***
Age	0-17	1.59 (1.57-1.61)***
	18-24	1.28 (1.27-1.29)***
	25-34	0.95 (0.94-0.96)***
	35-49	0.96 (0.95-0.97)***
Region	Africa & Middle East	2.43 (2.39-2.47)***
	Asia & Pacific	1.07 (1.05-0.08)***
	Australia	1.15 (1.13-1.17)***
	South & Central America	1.82 (1.80-1.84)***
	Europe	1.40 (1.39-1.41)

Table G.2 Multiple logistic regression for virus

Factor	Description	OR (95% CI)
Gender	Male	1.21 (1.19-1.23)***
Age	0-17	3.27 (3.14-3.40)***
	18-24	4.51 (4.37-4.66)***
	25-34	2.52 (2.44-2.60)***
	35-49	1.86 (1.80-1.93)***
Region	Africa & Middle East	14.67 (14.27-15.07)***
	Asia & Pacific	8.34 (8.20-8.58)***
	Australia	0.48 (0.43-0.54)***
	South & Central America	6.30 (6.13-6.47)***
	Europe	1.28 (1.24-1.31)***

Table G.3 Multiple logistic regression for cracks

Factor	Description	OR (95% CI)
Gender	Male	1.65 (1.62-1.67)***
Age	0-17	1.87 (1.81-1.92)
	18-24	2.86 (2.80-2.92)***
	25-34	2.41 (2.36-2.46)***
	35-49	1.68 (1.64-1.72)***
Region	Africa & Middle East	4.91 (4.79-5.03)***
	Asia & Pacific	3.30 (3.24-3.36)***
	Australia	1.37 (1.31-1.43)***
	South & Central America	4.81 (4.72-4.90)***
	Europe	2.16 (2.13-2.20)***

Table G.4 Multiple logistic regression for hack

Factor	Description	OR (95% CI)
Gender	Male	2.68 (2.47-2.91)***
Age	0-17	3.35 (2.30-3.74)***
	18-24	2.40 (2.18-2.63)***
	25-34	1.81 (1.65-1.99)
	35-49	1.58 (1.43-1.74)***
Region	Africa & Middle East	7.64 (6.94-8.41)***
	Asia & Pacific	2.39 (2.18-2.61)***
	Australia	1.35 (1.08-1.68)***
	South & Central America	4.97 (4.55-5.42)***
	Europe	2.90 (2.70-3.12)***

Table G.5 Multiple logistic regression for exploit

Factor	Description	OR (95% CI)
Gender	Male	1.48 (1.41-1.55)***
Age	0-17	0.67 (0.61-0.74)***
	18-24	0.88 (0.83-0.93)*
	25-34	1.04 (0.99-1.10)***
	35-49	1.05 (0.99-1.10)***
Region	Africa & Middle East	1.53 (1.37-1.70)
	Asia & Pacific	3.55 (3.38-3.73)***
	Australia	0.71 (0.60-0.85)***
	South & Central America	1.95 (1.81-2.09)***
	Europe	1.38 (1.31-1.45)**

Table G.6 Multiple logistic regression for rogue

Factor	Description	OR (95% CI)
Gender	Male	1.35 (1.30-1.40)***
Age	0-17	0.82 (0.76-0.89)
	18-24	0.66 (0.63-0.69)***
	25-34	0.69 (0.66-0.72)***
	35-49	0.74 (0.71-0.77)*
Region	Africa & Middle East	0.03 (0.02-0.04)***
	Asia & Pacific	0.007 (0.005-0.009)***
	Australia	0.27 (0.24-0.31)***
	South & Central America	0.009 (0.006-0.014)***
	Europe	0.05 (0.05-0.06)

Table G.7 Multiple logistic regression for infostealer

Factor	Description	OR (95% CI)
Gender	Male	1.52 (1.47-1.58)***
Age	0-17	2.64 (2.47-2.82)***
	18-24	2.52 (2.39-2.66)***
	25-34	1.92 (2.39-2.65)***
	35-49	1.48 (1.40-1.57)***
Region	Africa & Middle East	7.53 (7.04-8.06)***
	Asia & Pacific	7.85 (7.47-8.25)***
	Australia	1.59 (1.38-1.84)***
	South & Central America	23.73 (22.64-24.87)***
	Europe	2.04 (1.93-2.15)***

Table G.8 Multiple logistic regression for ransomware

Factor	Description	OR (95% CI)
Gender	Male	1.40 (1.32-1.49)***
Age	0-17	0.71 (0.62-0.80)***
	18-24	0.70 (0.65-0.75)***
	25-34	0.78 (0.72-0.84)
	35-49	0.93 (0.87-1.00)***
Region	Africa & Middle East	1.17 (1.02-1.33)***
	Asia & Pacific	0.47 (0.43-0.53)***
	Australia	0.90 (0.76-0.07)
	South & Central America	0.46 (0.40-0.53)***
	Europe	1.06 (1.00-1.13)***

Table G.9 Multiple logistic regression for bot

Factor	Description	OR (95% CI)
Gender	Male	1.53 (1.42-1.65)***
Age	0-17	0.98 (0.85-1.12)
	18-24	1.18 (1.09-1.29)***
	25-34	1.06 (0.97-1.16)
	35-49	1.07 (0.98-1.17)
Region	Africa & Middle East	4.94 (4.40-5.54)***
	Asia & Pacific	5.88 (5.45-6.34)***
	Australia	1.40 (1.13-1.76)***
	South & Central America	3.03 (2.73-3.36)***
	Europe	1.51 (1.39-1.64)***

Table G.10 Multiple logistic regression for rootkit

Factor	Description	OR (95% CI)
Gender	Male	1.65 (1.41-1.94)***
Age	0-17	1.10 (0.82-1.48)
	18-24	1.23 (1.02-1.47)
	25-34	1.43 (1.20-1.70)***
	35-49	1.29 (1.07-1.55)
Region	Africa & Middle East	0.71 (0.44-1.14)*
	Asia & Pacific	7.80 (6.83-8.92)***
	Australia	1.74 (1.20-2.51)*
	South & Central America	0.43 (0.28-0.66)***
	Europe	0.69 (0.57-0.83)***

## ANNEXE H ESTIMATES OF AVE AT 95%

Table H.1 Estimates of AVE at 95% by age group

	0-17		18-24		25-34		35-49		50+	
	AVE	(95% CI)	AVE	(95% CI)	AVE	(95% CI)	AVE	(95% CI)	AVE	(95% CI)
A	96.99	(96.46-97.43)	98.03	(97.84-98.20)	98.04	(97.86-98.21)	97.94	(97.76-98.10)	97.59	(97.39-97.79)
B	91.09	(90.18-91.81)	95.07	(94.69-95.43)	95.00	(94.58-95.39)	95.09	(94.77-95.39)	94.58	(94.26-94.89)
C	91.97	(91.07-92.77)	95.76	(95.50-95.99)	94.96	(94.64-95.25)	94.55	(94.24-94.84)	94.75	(94.48-95.01)
D	91.04	(90.18-91.81)	95.17	(94.93-95.41)	96.10	(95.88-96.31)	92.48	(91.97-92.97)	85.01	(83.56-86.34)
E	90.07	(89.25-90.82)	92.88	(94.05-94.53)	93.89	(93.61-94.16)	94.03	(93.72-94.32)	95.15	(94.78-95.49)
F	87.71	(87.08-88.30)	91.14	(90.87-91.40)	91.19	(90.93-91.44)	91.51	(91.30-91.71)	90.93	(90.73-91.13)
G	86.50	(85.91-87.06)	90.66	(90.44-90.88)	91.50	(91.28-91.71)	90.28	(90.02-90.53)	91.64	(91.36-91.90)
H	87.14	(86.34-87.89)	91.17	(90.86-91.47)	91.54	(91.23-91.84)	90.89	(90.51-91.26)	90.72	(90.26-91.16)
I	86.04	(85.61-86.47)	91.04	(90.88-91.19)	91.65	(91.49-91.81)	90.64	(90.44-90.84)	90.40	(90.15-90.65)
J	86.53	(86.01-87.04)	90.82	(90.61-91.03)	90.20	(89.97-90.42)	91.82	(91.67-92.04)	86.82	(86.58-87.06)

Table H.2 Estimates of AVE at 95% by region

	Africa & Middle East	Asia & Pacific	Australia	South & Central America	North America	Europe
	AVE (95% CI)	AVE(95% CI)	AVE(95% CI)	AVE(95% CI)	AVE(95% CI)	AVE(95% CI)
A	96.99 (96.46-97.43)	98.03(97.84-98.20)	98.04(97.86-98.21)	97.94(97.76-98.10)	97.59(97.39-97.79)	98.44(98.33-98.54)
B	91.09 (90.18-91.81)	95.07(94.69-95.43)	95.00(94.58-95.39)	95.09(94.77-95.39)	94.58(94.26-94.89)	92.54(91.75-93.25)
C	91.97 (91.07-92.77)	95.76(95.50-95.99)	94.96(94.64-95.25)	94.55(94.24-94.84)	94.75(94.48-95.01)	89.95(88.46-91.26)
D	91.04 (90.18-91.81)	95.17(94.93-95.41)	96.10(95.88-96.31)	92.48(91.97-92.97)	85.01(83.56-86.34)	85.93(85.21-86.62)
E	90.07 (89.25-90.82)	92.88(94.05-94.53)	93.89(93.61-94.16)	94.03(93.72-94.32)	95.15(94.78-95.49)	95.72(95.56-95.88)
F	87.71 (87.08-88.30)	91.14(90.87-91.40)	91.19(90.93-91.44)	91.51(91.30-91.71)	90.93(90.73-91.13)	93.24(93.08-93.40)
G	86.50 (85.91-87.06)	90.66(90.44-90.88)	91.50(91.28-91.71)	90.28(90.02-90.53)	91.64(91.36-91.90)	92.43(92.29-92.58)
H	87.14 (86.34-87.89)	91.17(90.86-91.47)	91.54(91.23-91.84)	90.89(90.51-91.26)	90.72(90.26-91.16)	93.48(93.33-93.63)
I	86.04 (85.61-86.47)	91.04(90.88-91.19)	91.65(91.49-91.81)	90.64(90.44-90.84)	90.40(90.15-90.65)	92.53(92.42-92.63)
J	86.53 (86.01-87.04)	90.82(90.61-91.03)	90.20(89.97-90.42)	91.82(91.67-92.04)	86.82(86.58-87.06)	91.21(91.06-91.36)

Table H.3 Estimates of AVE at 95% by HDI category

	Very high		High		Medium		Low	
	AVE	(95% CI)	AVE	(95% CI)	AVE	(95% CI)	AVE	(95% CI)
A	97.31	(97.17-97.44)	92.32	(99.24-99.38)	99.00	(98.86-99.13)	98.85	(98.37-99.18)
B	93.36	(93.13-93.58)	97.48	(96.94-97.92)	96.74	(95.95-97.37)	98.10	(94.90-99.29)
C	93.34	(93.15-93.52)	93.94	(93.15-94.64)	95.50	(94.30-96.45)	97.28	(94.73-98.59)
D	81.10	(80.21-81.94)	98.84	(98.78-98.89)	91.47	(90.79-92.10)	93.78	(91.80-95.28)
E	92.71	(92.46-92.96)	97.19	(97.09-97.29)	97.18	(96.94-97.40)	96.82	(96.05-97.44)
F	88.66	(88.50-88.81)	96.57	(96.39-96.74)	96.53	(96.22-96.81)	96.85	(96.04-97.48)
G	87.69	(87.51-87.88)	95.35	(95.21-95.49)	95.78	(95.54-96.00)	95.19	(94.48-95.82)
H	87.58	(87.31-87.85)	95.52	(95.32-95.71)	95.04	(94.63-95.42)	94.01	(92.66-95.12)
I	86.68	(86.51-86.84)	96.43	(96.36-96.49)	95.92	(95.74-96.09)	95.62	(95.13-96.06)
J	87.36	(87.22-87.50)	95.65	(95.53-95.77)	94.53	(94.29-94.75)	97.35	(96.82-97.79)